Is there a McGurk effect for tongue reading?

Olov Engwall

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

engwall@kth.se

Abstract

Previous studies on tongue reading, i.e., speech perception of degraded audio supported by animations of tongue movements have indicated that the support is weak initially and that subjects need training to learn to interpret the movements. This paper investigates if the learning is of the animation templates as such or if subjects learn to retrieve articulatory knowledge that they already have. Matching and conflicting animations of tongue movements were presented randomly together with the auditory speech signal at three different levels of noise in a consonant identification test. The average recognition rate over the three noise levels was significantly higher for the matched audiovisual condition than for the conflicting and the auditory only. Audiovisual integration effects were also found for conflicting stimuli. However, the visual modality is given much less weight in the perception than for a normal face view, and inter-subject differences in the use of visual information are large.

Index Terms: McGurk, audiovisual speech perception, augmented reality

1. Introduction

The McGurk effect [1] is a well-known and well-studied phenomenon in normal audiovisual speech perception, indicating that conflicting auditory and visual speech stimuli are integrated automatically and unconsciously by the listeners. Often, simultaneous presentation of an auditory speech stimulus A (e.g., [b]) and a visual speech stimulus V (e.g., [g]) results in the perception of an intermediate category C (e.g., [d]) that is the most compatible with the combined information provided by A and V. ‘Normal’ here refers to the fact that the visual articulatory information is linked to facial movements, of e.g., the lips, which the listeners are accustomed to seeing. It has however been shown that the effect appears also for computer-animated faces [2], for highly reduced face images [3] or when the listeners are unaware that they are looking at a face [3]. Young infants also perceive the effect [4], as do listeners who touch the speaker’s face rather than seeing it [5]. This might be interpreted as evidence that the integration of different modalities does not need to be learned actively and that visual and sensory information about speech may be encoded together with the acoustic information.

However, studies [6, 7, 8, 9] using an augmented reality (AR) display of the face, such as the one in Figure 1, to display tongue movements have indicated that tongue reading capabilities are weak initially, thus contradicting the hypothesis that speech perception and production and the two audiovisual modalities are so closely connected that unfamiliar visual articulatory information could be directly decoded.

Grauwinkel et al. [6] performed an audiovisual consonant recognition task in noise and showed that the group that had received training on tongue reading performed significantly better than both the group that had not received any training and the one that was shown a normal side view of the face, without tongue movements visible. The training was in the form of an instruction video that explained the movement of the articulators for all consonants in all vowel contexts in a side view of the face with semi-transparent skin. Badin et al. [7] performed a similar audiovisual perception test of VCV words, where one group started from clear condition and was presented the stimuli in increasing levels of noise while the other started from muted condition, from which the signal-to-noise ratio (SNR) level was increased. The first group of subjects, who received implicit training of the relationship between acoustics and tongue movements in the initial clear condition, had as a group significantly higher recognitions scores in noise than the group that instead started in muted condition. Engwall & Wik [8, 9] studied audiovisual perception of acoustically degraded sentences and VCV words. The first study showed that some sentences containing phonetic features that are difficult to see in a normal face view were better perceived if an AR side view displaying tongue movements was added to a normal front face view. Over all, the recognition scores were however not higher than if only the normal face view was presented. The second study showed that both VCV words and sentences were perceived significantly better if accompanied by tongue animations in a AR side view, than if only the auditory signal was presented. Further investigated the difference in recognition score between tongue animations generated from measurements and from rule-based synthesis. For VCV words, synthetically generated movements resulted in higher recognition rates, while for sentences, the movements recreated from Electromagnetic Articulography (for the tongue) and motion capture (for the face) data were better. In both studies, a short familiarization phase, in which tongue animations were displayed together with clean audio, preceded the test.

Figure 1: An augmented reality (AR) view of the talking head displaying tongue movements.
The four above studies suggest that at least a short, explicit or implicit, training may be required if listeners are to benefit from animations of tongue movements in their speech perception. In order to investigate whether this learning is that the actual mapping between the displayed tongue movements and the acoustics is being established during the experiment (i.e., the animated movements have no articulatory significance to the subjects and the animations are arbitrary icons) or that it is the articulatory awareness that is increased (i.e., subjects learn that the unfamiliar animations may contain relevant information and use articulatory knowledge that they already had prior to the test), this paper investigates audiovisual integration with inconsistent visual information. This signifies that an auditory signal in noise is presented together with randomly alternating matching or conflicting tongue movements, which makes iconic learning during the test impossible. As a consequence, if the mapping is learned during the test, then recognition scores will be no better with matched tongue movements than with conflicting. On the other hand, if the subjects do make use of an already established knowledge about the acoustic-articulatory relation, then the matched audiovisual stimuli will result in higher recognition scores and conflicting stimuli should give rise to interpretation effects similar to those for conflicting face images, as first described by McGurk and McDonald [1].

The test is focused on a set of hypotheses related to speech perception when contradictory animations of tongue movements are presented together with the acoustic signal:

(H1) In clear condition, the acoustic signal dominates the perception entirely, regardless of if the animation matches or not, and no McGurk effect occurs, since the information provided by the visual modality is much weaker.

(H2) When the acoustic signal is degraded by noise, subjects will start making use of the visual modality in their interpretation. Matching movements should therefore result in higher identification scores and a McGurk effect will appear for conflicting audiovisual stimuli.

(H3) Since tongue movements are visually less familiar than face movements, the McGurk effect will be weaker than previously reported for conflicting face videos or animations.

(H4) Just as lip-reading capabilities differ between individuals, subjects differ in tongue reading skills, and this will influence their response to contradictory stimuli, so that better tongue readers are more sensitive to the audiovisual conflict.

2. Test conditions

The experiments conducted to investigate the above hypotheses were performed using the side-view of a talking head with transparent skin at the cheek, as shown in Figure 1. The side-view was chosen, since the differences between the places of articulation for the tongue are clearer than in a front view, which is normally used for McGurk tests with face images. Furthermore, a side-view was used in previous studies on tongue reading [6, 7, 8, 9]. 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 for the face and dorsum raise, body raise, tip raise and tip advance for the tongue. As will be explained in Section 2.1 below, the five parameters for the lips (i.e., lip rounding, upper and lower lip raise and retraction) were not used in the creation of the animations, to avoid that lip reading influences the results.

The tongue and jaw models are based on a statistical analysis of Magnetic Resonance Imaging (MRI) data of a Swedish subject producing static vowels and consonants in VCV context [10]. The position and shape of the models for each phoneme are hence derived from real articulations, which should ensure that they present relevant articulatory information. The movements between these static configurations were created using rule-based visual speech synthesis [11]. Rule-based synthesis was chosen over resynthesis from EMA data, since it allows for automatic creation of time-aligned conflicting audiovisual stimuli, whereas resynthesis from data would require a more complex time warping to align it with unmatched animations. A previous study [9] further indicated that rule-based synthesis gave better recognition results for VCV words than resynthesis.

2.1. Stimuli and Conditions

The 192 stimuli consisted of 24 different VCV words spoken by a female speaker of Swedish, presented in four blocks with different levels of white noise and three different audiovisual conditions. All subjects were presented three noisy blocks in order of decreasing signal-to-noise ratio (SNR): first +3dB, then -6dB and last -9dB, followed by a final Clear condition (no noise). At each noise level, the same set of 48 VCV words was played, in random order between noise levels and audiovisual conditions, but in the same order for all subjects. Of the 48 words, 24 were played in acoustic only condition (AO), to provide the recognition base-line. These 24 words were all combinations of the consonants C=[p, b, t, d, k, g] and the vowels V=[a, i, u] in symmetric VCV words. The same 24 words were also played in audiovisual condition, 12 each with matching (AVM) and conflicting (AVC) animations of the tongue movements. The animations were rotated so that the auditory signal for each of the stop phonemes [p, b, t, d, k, g] was presented together with visual stimuli containing tongue movements related to each of the three places of articulation (bilabial, alveolar, velar). [i, u] constituted a separate conflict pair, i.e., the non-matching condition for [i] was visual [v] and vice versa.

In order to avoid that the exact same VCV word was presented to the same subject in both matching and conflicting condition, the subjects were divided into two groups, which were presented the stimuli in opposite matching conditions. The combinations of vowel contexts, stop consonant voicing and place of articulation were distributed evenly between the two groups. This means that if [ap:a, Id:I, uk:u] constituted a separate conflict pair, i.e., the non-matching condition for [i] was visual [v] and vice versa.

The 3SNR levels were selected to provide conditions where either the acoustic or visual information would be the strongest (SNR=+3dB and -9dB, respectively) and one in which the combination would be important (SNR=-6dB).
2.2. Subjects

18 subjects, without any known hearing impairment, participated in the test, 13 male and 5 female (aged 21-31 years). Of these, four were native speakers, five had a European native language (German, English, Serbian, Greek), five a non-European Indo-European native language (Persian, Urdu, Bangla) and four a native language belonging to another language family than Swedish (Chinese, Korean, Thai, Tamil). The reason for including non-native speakers in the test is that the main targeted application for the augmented reality talking head display is for pronunciation training in second language learning [12], where it is used to clarify articulatory differences in the tongue shape and position. It is hence of interest to start investigating if the subject’s language background influences the ability to interpret animations of tongue movements or if the audiovisual integration is universal, regardless of the similarity between the first and target languages.

2.3. Experimental set-up

The acoustic signal was presented over high quality headphones and the articulatory animations were displayed on a 21” flat screen. Each stimulus was presented once, without repetition, and subjects answered in a forced choice setting, by selecting the on-screen button labeled with the consonant corresponding (the most) to what they heard.

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. Since one hypothesis was that the auditory information would completely dominate in the Clean condition, this set was presented last, in order to reduce the risk that subjects learned that the visual information could not be trusted, since the discrepancy would be more noticeable with clean speech.

The visual stimuli were displayed in a side-view of the face for all three conditions. For AVC and AVM, the augmented reality view of the face (c.f. Figure 1) was shown with tongue and jaw movements, whereas a normal, outside view of the face without any movements was shown for AO.

The test started with a familiarization phase, in which 9 VCV words with \( C=\{m, n, r\} \) and \( V=\{a, i, u\} \) were presented in matching audiovisual condition. Each of the nasal stops was presented once at each of the three noise levels Clear, -6dB and -9dB, in order of decreasing SNR. Since no feedback was given on the subject’s replies and the familiarization stimuli were not included in the test, this should not be considered as training for the following test.

When the subjects had completed the test, they were asked by the presentation interface to briefly describe if and how they had used the animations of the tongue to interpret the stimuli. These informal comments were used to investigate if any trends could be found between the subject’s claimed awareness of tongue movements and the recognition results.

The entire experiment, including familiarization, test of the 192 VCV words and the entry of informal comments lasted about 15 minutes per subject.

3. Results

The consonant accuracy rates were counted automatically by the test program. Since the aim of the study was to investigate audiovisual integration with respect to tongue articulation, and in particular since some subjects were from language backgrounds where no voiced-voiceless distinction is made, responses were grouped as \{\( p/b \), \( t/d \), \( k/g \), \( l/j \)\} and errors of voicing were disregarded. This means that the chance level is 20%.

For each noise level and presentation condition, two recognition scores were calculated, one for all consonants and one for \{\( t, d, k, g, l \)\}, excluding the labial consonants \( p, b, v \), since the AVM animation for the latter constitutes a special case, as described in Section 2.1. For the AVC condition, recognition accuracy was defined with respect to the auditory stimuli.

Figure 2 and Table 1 summarize the general results, in terms of accuracy levels (Figure 2(a)) and differences compared to the acoustic only condition (Figure 2(b)) for different SNR levels, and mean identification scores for the two factors SNR level and presentation condition.

Table 1: Mean accuracy rate \( m \) and standard deviation \( \sigma \) as a function of signal-to-noise ratio and presentation condition.

<table>
<thead>
<tr>
<th>SNR</th>
<th>( m )</th>
<th>( \sigma )</th>
<th>Cond.</th>
<th>( m )</th>
<th>( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>+3dB</td>
<td>0.53</td>
<td>0.03</td>
<td>AVM</td>
<td>0.55</td>
<td>0.04</td>
</tr>
<tr>
<td>-6dB</td>
<td>0.48</td>
<td>0.04</td>
<td>AO</td>
<td>0.45</td>
<td>0.02</td>
</tr>
<tr>
<td>-9dB</td>
<td>0.42</td>
<td>0.03</td>
<td>AVC</td>
<td>0.44</td>
<td>0.04</td>
</tr>
</tbody>
</table>
and presentation condition (Table 1). Note that while individual subjects were not presented the exact same stimuli in both AVM and AVC, all subjects were presented all stimuli in AO, thus making a direct comparison between AVM and AO, and between AVC and AO, possible. The results were therefore analyzed based on the difference between each of the matching/conflicting conditions and the acoustic only baseline. Since the combinations of vowel context, voicing and place of articulation were rotated and distributed evenly between AVM and AVC for each subject, comparisons between these two conditions should also be valid, even if they should be made with more caution, since they relate to similar, but not identical, stimuli (differing in vowel context or voicing).

A two-factor ANOVA on SNR level (excluding the Clean condition) and presentation condition shows that both are significant, but the interaction is not. The AVM recognition rate is significantly better than the AO and AVC rates at p<0.005, while there is no significant difference between the AO and AVC rates. When presentation condition is tested for significance separately at each SNR level, only the differences between A VM-AO and A VM-A VC at SNR=-6dB are significant (at p<0.005 using a single factor ANOVA).

For the clean condition, the recognition is almost perfect, regardless of visual stimuli type, since the auditory signal totally dominates the perception. At +3dB, the benefit of the visual information starts appearing, but as a group, the subjects do not perform significantly better with AVM than in acoustic only condition, and in particular, AVC does not degrade the perception further from the AO level. Noteworthy is that the recognition scores for both AVM and AVC are lower when the labial consonants are included, indicating the subjects were more sensible to both lacking (for AVM) and conflicting visual information related to the lips than to information on which part of the tongue (tip or dorsum) that is used.

3.1. McGurk effects

Figure 3 summarizes the subjects’ responses to different audiovisual stimuli combinations at SNR=-6dB, at which the clearest effects appear. The graph shows that there are indeed both expected integration of coherent audiovisual stimuli and phenomena similar to the McGurk effect, even if they are much less universal between subjects and weaker than for lip reading. Seven cases have been highlighted in Figure 3 (all percentages of response changes are relative to the AO baseline):

1. For auditory [p/b], the matched condition shows no labial movement and this leads to a decrease in [p/b] responses and an increase in [k/g] responses compared to the acoustic only case (this can be considered to be a case of "normal" McGurk effect, related to the lips).

2. However, for auditory [p/b] in combination with visual [k/g], the [p/b] responses increase, while the [t/d] and [k/g] responses decrease by 50%. This hence differs from the traditional McGurk effect, in which [t/d] would be the preferred reply with a normal face view. The effect is even stronger at SNR=-9dB.

3. When auditory [t/d] is combined with visual [k/g], [p/b] responses increase by 100% and [l] responses disappear. That is, some subjects rule out [k/g] based on the auditory signal and [t/d, l] based on the visual and respond with the least conflicting, [p/b]. It should be noted that this does not happen at SNR=-9dB, where [k/g] and [v] responses instead increase and [p/b] responses disappear.

4. When auditory [k/g] is combined with visual [p/b], [k/g] responses decrease by 40% and [t/d] responses increase by 40%. This is thus the reverse of the normal McGurk effect. At SNR=-9dB, both [l] and [t/d] responses increase.

5. The matched [k/g] condition illustrates the benefits of coherent audiovisual information on the tongue articulation.

6. Auditory [v] in combination with visual [l] results in a 50% increase of [t/d] responses and a 60% decrease in [p/b]
responses, compared to AO. (7) In the “matched” [v] condition, the [t, d] responses decrease by 60%.

It can also be noted that the auditory information for [l] is so salient that it dominates the perception almost entirely, even in non matched condition at the lowest SNR=-9dB.

3.2. Subject differences

The accuracy rates for individual subjects, shown in Figure 4, indicate that there are large inter-subject differences in tongue reading ability. While the AO score at SNR=-9dB is rather similar between subjects (mean $m_{AO}=0.34$, standard deviation $\sigma=0.09$), the AVM score at SNR=-9dB ranges from 0.14 to 0.88 ($m_{AVM}=0.48$, $\sigma=0.20$). This hence means that subjects who achieve higher AVM scores at low SNR levels do so because they make more use of the visual information, not because they are better at discriminating the auditory signal in noise. The AVM-AO difference is positive for about 2/3 of all the subject-SNR combinations (excluding the Clean condition). It can be noted that several subjects require a higher SNR to be able to successfully integrate audiovisual information, with peak performance in AVM compared to AO at SNR=-3dB (subjects 5, 6, 7, 9) or at SNR=-6dB (subjects 1, 10, 18).

A two-way ANOVA with subjects and presentation conditions as factors indicates that both factors and their interaction are significant at $p<0.001$. The average AVM score per subject over all noisy SNR levels ranges from $m_{AVM}=0.83$ (standard deviation $\sigma=0.04$, $m_{AO}=0.48$, $m_{AVC}=0.33$) for the “best” tongue reader down to $m_{AVM}=0.29$ ($\sigma=0.04$, $m_{AO}=0.30$, $m_{AVC}=0.33$) for the subject with the lowest score. In general, subjects with higher AVM scores have not only a larger positive AVM-AO difference, but also a larger negative AVM-AO difference. The average AVM score per subject over all noisy SNR levels ranges from $m_{AVC}=0.17$ ($\sigma=0.04$, $m_{AO}=0.4$, $m_{AVM}=0.71$) up to $m_{AVC}=0.67$ ($\sigma=0.005$, $m_{AVM}=0.67$, $m_{AO}=0.70$). Since the subjects who had the lowest $m_{AVM}$ and the highest $m_{AVC}$ achieved approximately the same recognition rates over all three presentation conditions, they may in fact not have been considering the visual signal at all. The AVM scores are even on average 10% higher than for AO for the 8 “weak” tongue readers over the three noisy conditions, compared to 7% lower for the 10 other subjects.

Earlier studies [8, 9] have also shown that perception scores differ substantially, even if all subjects are native, but in the current study language background is also a contributing factor, as discussed further in Section 3.2.2. Of the seven subjects with the highest scores in Figure 4, three are native (subjects 13, 15, 16) and three are European (subjects 12, 17, 18), indicating that some inter-subject differences may be due to familiarity with the test language. However, one of the subjects with the lowest score is native (subject 3) and one has a European first language (subject 2). Individual differences in how the subjects tried to use the animations may hence be as important.

3.2.1. Informal comments

Interpretations of the subjects’ statements on if and how the tongue movements had been useful run the risk of becoming anecdotal, but some trends are nevertheless observable when considering the comments in relation to the AVM and AVC scores. Note that the subjects did not know their recognition accuracy when giving these comments.

(1) Subjects with high AVM accuracy rates and/or large AVM-AVC differences stated that they were actively considering the tongue (Subject 18: “I used the tongue whenever the sound was not audible.”), Subject 17: “In difficult cases, I tried to redo the tongue movement and deduce the consonant.”, Subject 16: “When the sound was bad I looked at the tongue, to see where the top of it or the back of it were.”, Subject 10: “I found it confusing sometimes. But in some cases, I could use the place of articulation to be sure.”, Subject 9: “tongue helps a bit to guess the pronunciation, especially [f] there is bigger noise.”.

(2) Subjects with small AVM-AVC differences may have actively refrained from looking at the tongue animations, because they did not trust them (Subject 15: “The tongue was very confusing, I tried to say the sound myself and see if my tongue matched. I thought that the tongue sometimes matched and sometimes didn’t.”, Subject 7: “Difficult to use the tongue without lips movement.”, Subject 13: “The tongue was irritating with high noise because then I “saw” one sound and heard another one.”).

(3) Subjects with low AVM accuracy rates and/or negative AVM-AVC differences seem to have either refrained from considering the visual stimuli (e.g., Subject 8: “carefully listening to the sound of pronunciation, not too much reference [to] the video on screen.”), Subject 3: “I tried to use it sometimes but it is hard to try to figure out how the tongue actually moves when you talk so mostly I just listened”) or were merely distracted by seeing the unfamiliar movements (Subject 1: “It was difficult for me to understand, I tried to feel my tongue, it was hard.”, Subject 4 stated that he was only guessing for the noisy conditions, but claimed to have learned in Clear condition and suggested that he would have performed better if the set order had been reversed.)

3.2.2. Differences due to language background

Since the number of subjects per language group is too small to make any statistical comparisons, the accuracy rates grouped by similarity between the subjects’ first language (L1) and Swedish cannot answer the question if the similarity influenced the ability to tongue read the VCV stimuli. Figure 5 do however illustrate two possible trends, namely that: 1) The similarity could indeed be an influencing factor (the mean AVM accuracy
Figure 5: AVM accuracy rates (curves) and differences $\Delta=AVM-AVC$ in accuracy rates (bars) for different language groups for $[t,d]$, $[k,g]$, $I$, $v$. Chance level is 20%.

rate over the three noisy conditions is $m=0.62$, $\sigma=0.21$ for the Swedish and European subjects, and $m=0.48$, $\sigma=0.18$ for the others. 2) Non-native subjects may start to use visual cues at higher SNR levels than native speakers, who initially rely more on the auditory cues (shown by the delayed increase in $\Delta$ for native subjects).

4. Discussion and conclusions

The results of this study indicate that the auditory stimuli are too dominating compared to visual stimuli of tongue movements for the McGurk effect to appear in clear auditory condition (H1). With degraded audio, the visual information is given more importance by the subjects and McGurk-like effects appear (H2), but they are not as strong as with conflicting facial information and not universal between subjects (H3). It was also observed that some subjects were very strong tongue readers, achieving significantly higher scores with matching tongue movements than with conflicting, while others did not get any support at all from the matching visual information (H4). The subjects in the first group hence seemed to be aware of their tongue movements, while the latter either choose not to use the visual information or tried to learn the mapping during this test.

A natural follow-up study to test if tongue reading abilities are learned or already established would be to use an initial acoustically clear condition as a training phase and compare three different training-test conditions. One group of subjects would be trained on coherent audiovisual data (I) and two groups on incoherent (II & III). All groups would be tested in noisy condition on both coherent and incoherent data, to investigate the importance of coherence between the acoustic and visual information and of consistency between training and test. Group II would be trained on randomly incoherent stimuli, while group III would be trained on conflicting, but consistent audiovisual combinations. That is, they would always be shown e.g., the articulatory movements for $[k,g]$ together with the acoustics for $[t,d]$ and vice versa. If subjects actually learn the mapping during the initial phase without any articulatory relevance, then group III would be as successful as group I for audiovisual combinations that are consistent between training and test, whereas group II would score lower. On the other hand, if the audiovisual relationship is already established, all groups would score comparatively higher with matching test data and group I would reach a higher score than group III for test stimuli that are consistent with the training.

There were indications that the subjects’ first language influences their ability to interpret tongue movements in the target language. This should be investigated further, since the ability to understand animations of tongue movements are crucial for computer-assisted pronunciation training feedback relying on displaying intra-oral articulatory differences. As a first step, the test should be re-run with native subjects only, to investigate if the inter-subject differences are smaller for a more homogeneous group.

With the above described weakness of tongue animations as compared to face images and the need for further investigations in mind, a main finding of this study is nevertheless positive with respect to tongue reading: Since the subjects as a group were significantly better at identifying consonants with matching tongue movements, despite the setup that made learning of audiovisual mapping during the test impossible, and since audiovisual integration of conflicting stimuli was observed, the subjects must have had an already established awareness of the relation between acoustics and tongue articulations, which they can use to support speech perception.

5. Acknowledgments

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

6. References