Acoustic and visual phonetic features in the McGurk effect – an audiovisual speech illusion

Kaisa Tiippana [1], *Mikko Tiainen* [1], *Lari Vainio* [1], *Martti Vainio* [1]

[1] Institute of Behavioural Sciences, University of Helsinki, Finland

kaisa.tiippana@helsinki.fi, mikko.o.tiainen@helsinki.fi, lari.vainio@helsinki.fi, martti.vainio@helsinki.fi

Abstract

Information from the acoustic speech signal and the talking face is integrated into a unified percept. This is demonstrated in the McGurk effect, in which discrepant visual articulation changes the auditory perception of a consonant. We studied acoustic (A) and visual (V) phonetic features that contribute to audiovisual speech perception by measuring the McGurk effect in two vowel contexts, [a] and [e], at various levels of acoustic noise. The McGurk stimuli consisted of an acoustic [p] presented with a visual [k]. This combination is generally heard as [t] (called fusion) or as [k] (visually dominant percept). The stimulus A[apa]V[aka] was most often heard as [aka], and these percepts increased with noise level. The stimulus A[epe]V[eke] was heard mostly as a fusion [ete], but in high noise also as [eke]. A phonetic analysis showed that, in [e] context, A[p] and V[k] stimulus features were close to those of [t], explaining why fusions were frequent. In [a] context, the visual stimulus had clear features of [k], while the features of the acoustic component were less distinctive, resulting in visual dominance particularly in noise. These results show how audiovisual integration depends on the features of acoustic and visual speech.

Index Terms: speech perception, multisensory, audiovisual, integration, McGurk effect

1. Introduction

Integration of the acoustic speech signal with the visible articulatory gestures on the talker's face normally enhances speech perception, particularly in noisy conditions [1]. The auditory and visual features that contribute to the integrated speech percept are still poorly known. A widely used tool to study audiovisual integration of speech is an illusion called the McGurk effect [2]. In the classical McGurk effect, an auditory consonant [b] that is presented with a face articulating [g], i.e. A[b]V[g] stimulus, is generally heard as a fusion [d] [2]. The fusion effect has raised the most interest since the percept differs from the acoustic and visual stimulus components. Sometimes this kind of a stimulus may also be heard as [g], producing a visually dominant auditory percept. Here we refer to the McGurk effect as an auditory percept differing from that produced by the acoustic signal when presented alone, be it a fusion or a visually dominant percept (see [3] for a discussion of this definition). The strength of the McGurk effect reflects the strength of audiovisual integration. That is, the fewer responses according to the acoustic stimulus are given, the stronger the visual influence and thus the McGurk effect, and the stronger the integration. The McGurk effect occurs even when the acoustic signal presented alone is clear and perfectly recognized.

Two different instances of unvoiced McGurk A[p]V[k] stimuli have been used in our previous studies [4,5]. Voiceless stop consonants were used because [b] and [g] are not native Finnish consonants. The stimuli were vowel-consonant-vowel nonsense words. The vowel was [a] in one study [4] and [e] in the other study [5]. The [a] context stimulus was heard as a visually dominant [k], particularly at high levels of acoustic noise [4]. The [e] context stimulus was heard predominantly as a fusion [t], but here the McGurk effect was measured only without noise [5].

In the current study, we addressed the question why the same consonants in these McGurk A[p]V[k] stimuli produce different percepts when they are presented in different vowel contexts, [a] and [e]. We also studied the effect of acoustic noise on speech perception in both vowel contexts. Our hypothesis was that when noise is added to the [e]-context stimulus, the proportion of T responses should decrease, while the proportion of K responses should increase. This is because visual influence gets stronger when the acoustic signal becomes less reliable [6] as its phonetically relevant features are masked. We also investigated the stimulus features contributing to the differences in perception.

2. Methods

2.1. Participants

Twenty Finnish adults with a mean age of 23 years (range 19-32 years) served as participants. All reported normal hearing and vision, and no language-related problems.

2.2. Stimuli

The stimuli were the same as used in our previous studies [4,5]. Video recordings of two female talkers uttering meaningless VCV words were edited to produce auditory, visual and McGurk stimuli. The vowel was [a] for one talker and [e] for the other, and the consonants were [p], [t] and [k] for both. Of the main interest were the two McGurk stimuli: A[apa]V[aka] and A[epe]V[eke], which were created by replacing the original soundtrack of the [k] video with the sound file with [p] by aligning the onsets of the waveforms. Original, congruent audiovisual stimuli were also included as fillers, but the results are irrelevant and not reported here.

The acoustic stimuli were presented in white noise at signal-to-noise ratios (SNR) of -12, -6, 0 and +6 dB, and without noise at 46 dB(A). They were delivered via headphones. The visual stimuli were presented as bitmap sequences in the centre of a LCD screen.

2.3. Procedure

The [a] and [e] context stimuli were presented in separate blocks in counterbalanced order across participants. The McGurk, auditory and visual stimuli were presented ten times in randomized order within a block, and the filler stimuli only twice. After each stimulus, the participant pressed a key on a computer keyboard according to what they heard for the

acoustic and audiovisual stimuli, or lip-read the visual-only stimuli. The results were analyzed for the most common responses, which corresponded to the presented stimuli: K, P and T.

3. Results

3.1. The effect of noise on the McGurk effect

The results for the [a] context stimuli replicated our previous findings [4]. The participants heard the A[apa]V[aka] stimulus mostly as [aka], and the proportion of K responses increased as the acoustic noise level increased [$F(4,76)=30.3$, $p<0.001$; repeated-measures ANOVA, Greenhouse-Geisser corrected p, uncorrected df reported] (Fig. 1a). There was a corresponding decrease in P [$F(4,76)=6.63$, $p=0.007$] and T [$F(4,76)=5.28$, $p=0.008$] responses with decreasing SNR. Thus, the visual dominance increased when the acoustic signal got poorer.

The [e] context stimulus presented without noise also replicated our previous finding [5] since the A[epe]V[eke] stimulus was heard mostly as [ete] (Fig. 2a). In addition, when the acoustic SNR got poorer, the proportion of these fusion T responses first increased, but then decreased. There was a corresponding increase in K responses at the poorest SNRs.

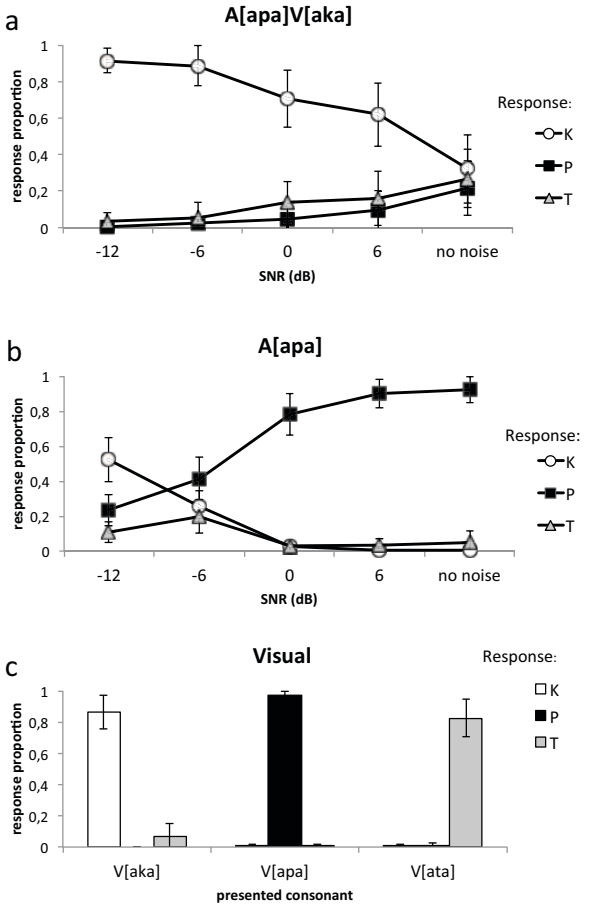

Figure 1: *Response proportions for the McGurk stimulus (a) and the acoustic stimulus [p] (b) as a function of acoustic noise level, as well as for the visual stimuli (c) in [a] context.*

The proportion of P responses dropped to near zero in noisy conditions. The shift towards an increased visual weighting as the acoustic signal became less reliable was even clearer for the [e] than [a] context stimuli since for the former there was a gradual shift from some P responses to mostly T responses into more K responses as the SNR got weaker, producing a main effect of SNR [K: $F(4,76)=35.5$, $p<0.001$; P: $F(4,76)=8.61$, $p=0.005$; T: $F(4,76)=15.1$, $p<0.001$].

3.2. The effect of noise on auditory speech perception

For the [a] context stimuli, the proportion of correct responses for auditory [p] decreased as the SNR got poorer, and there were increasingly more other responses [main effect of SNR for K: $F(4,76)=51.0$, $p<0.001$; P: $F(4,76)=57.5$, $p<0.001$; T: $F(4,76)=7.53$, $p=0.001$], so that the response distribution approached random guessing at the lowest SNRs (Fig. 1b), as could be expected.

For the [e] context stimuli, the proportion of correct responses for auditory [p] also decreased as the SNR got poorer (Fig. 2b). However, the transition in other responses was not smooth towards a random distribution. Instead, T responses first increased and then decreased with decreasing SNR. The main effect of SNR was significant for all responses [K: $F(4,76)=27.3$, $p<0.001$; P: $F(4,76)=98.4$, $p<0.001$; T: $F(4,76)=35.7$, $p<0.001$].

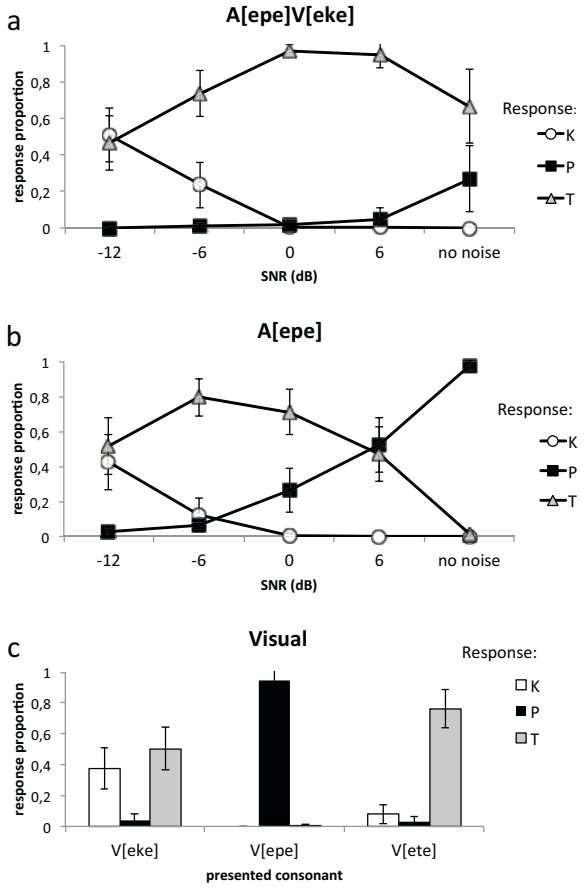

Figure 2: *Response proportions for the McGurk stimulus (a) and the acoustic stimulus [p] (b) as a function of acoustic noise level, as well as for the visual stimuli (c) in [e] context.*

3.3. Visual speech perception

For the lip-reading task in the [a] context, visual [p] was recognized more accurately than [k] (p=0.047, Bonferroni corrected pairwise comparison) and [t] (p=0.025), which did not differ from each other (Fig. 1c).

In the [e] context, the recognition of visual [p] was the best, followed by [t], meanwhile [k] was poorly recognized (all pairwise comparisons p<0.05) (Fig. 2c). Visual [k] was frequently lip-read as [t].

3.4. Phonetic analysis of the acoustic speech stimuli

To investigate the underlying reasons for the differences in auditory and audiovisual speech perception between the [a] and [e] context stimuli, spectrograms of the acoustic stimuli were created, and the first (F1) and second (F2) formants were estimated using Praat[6] (Fig. 3). This phonetic analysis revealed vowel-dependent differences in F2 transitions and release bursts between the consonants, which could partially account for the perceptual effects.

In [a] context, the F2 transition from the consonant towards the vowel was rising for [p], but falling for [k] and [t]. Thus, the transitions differed clearly between [p] and [k], as well as between [p] and [t], with a smaller difference between [k] and [t]. Adding noise to these stimuli then gradually made the [p] roughly equally confusable with both [k] and [t]. For the audiovisual stimuli, the weaker reliability of the acoustic signal resulted in a stronger visual influence seen as more K-responses at higher levels of noise.

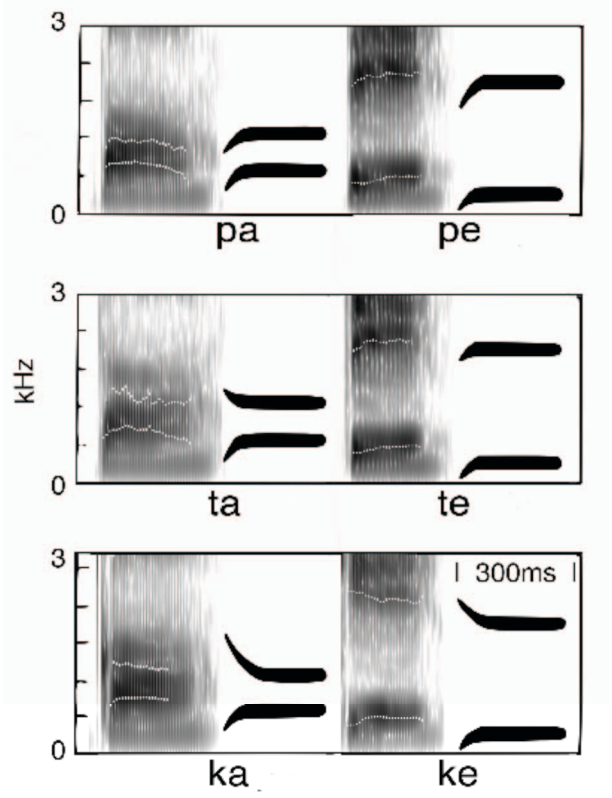

Figure 3: *Spectrograms of the acoustic stimuli. The crucial latter CV syllable is shown, together with F1 and F2 fits (dotted lines) and their graphical depictions [8] (black lines on the right).*

The release burst was weak for [p] and [t]. For [k] it was stronger, peaking around 1.4 kHz, making it harder to mask with the wide-spectrum noise.

In [e] context, the F2 transition was rising for [p] and [t], but falling for [k]. Thus, the transitions were similar for [p] and [t], but clearly different for [k]. Furthermore, the release burst was weak for [p], wide-spectrum for [t] and peaked around 3 kHz for [k]. When white, i.e. wide-spectrum noise was added, the burst for [p] started to resemble that of [t]. For these reasons, [p] was confused with [t] in noise. This effect first got stronger with increasing noise level, seen as an increase in T-responses, but at the poorest SNR it started to decrease for both auditory and audiovisual conditions.

3.5. Phonetic analysis of the visual speech stimuli

The visual speech stimuli were analysed by inspecting the videos frame by frame during the articulatory movements. Attention was paid to the extent of mouth opening, visibility of the tongue and other articulators, and particularly the features distinguishing [k] from [t].

The visual stimuli were more distinctive in [a] vowel context than in [e] vowel context (Fig. 4). This was mainly because the mouth was more open when [a] rather than [e] was uttered. Furthermore, the visual quality was better for the [a] context stimuli since the inside of the mouth was lit well, improving the visibility of the tongue; and the video images were sharper, enhancing visual detail. Consequently, the position of the tongue was more visible for [a] context stimuli.

At the moment just before the consonant release, the tongue forms a constriction at the velum for [k], and at the alveolar ridge for [t]. In [a] vowel context this difference was clearly visible so that the tongue was evidently far back in the mouth for [k], and in the front behind the teeth for [t]. Instead, in [e] vowel context the [k] and [t] articulations looked very similar.

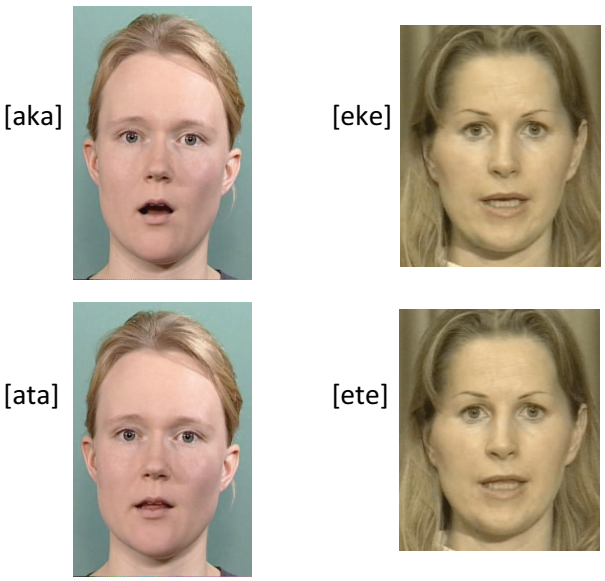

Figure 4: *Video frames from the visual stimuli just before the consonant release: top row [k], bottom row [t]; left side [a] context. right side [e] context.*

4. Discussion

We found that a McGurk stimulus consisting of an acoustic [p] and a visual [k] was mostly heard as a fusion [t] in [e] vowel context, while in [a] vowel context it was mostly heard according to the visual component [k]. The effect was modulated by the level of acoustic noise, so that the visually dominant responses increased with decreasing SNR.

The strength of the McGurk effect was not affected by vowel context, but instead it was altered qualitatively – i.e. there was no change in the proportion of responses corresponding to the acoustic component, but in the type of visually influenced responses – in agreement with previous studies [9-13]. For example, more audiovisual fusion percepts have been found in [i] than [a] context in Japanese, accompanied by more confusions between visual [d] and [g] [9]. Our results correspond to this in that a closed vowel produced more audiovisual fusion percepts than an open vowel, together with more visual confusions between an alveolar and a velar stop consonant.

The phonetic analyses of our acoustic and visual stimuli provided explanations for the perceptual differences.

The acoustic analysis showed that the F2 transitions were similar for [p] and [t] in [e] context, but different from each other in [a] context. Thus, [p] and [t] were acoustically more similar in [e] than [a] context. Additionally, the release burst of [t] was wide-spectrum, resembling white noise in [e] context. These acoustic features led [epe] to be confused with [ete] in white noise.

The visual analysis showed that there were more distinguishing features between the visual articulations of [k] and [t] in [a] than [e] vowel context. In other words, [k] and [t] were more confusable to each other in [e] than [a] vowel context, as reflected also in lip-reading performance, which is known to be more accurate with open than close vowels [14]. This, in turn, was reflected into the McGurk effect so that there were more fusion [t] percepts in [e] than [a] context.

For the McGurk stimulus in [a] context, the visual weighting in the integrated speech percept was very strong because of the clear and reliable visual features, resulting in mostly K responses. In [e] context both stimulus components, the acoustic [p] and the visual in [k] shared features with [t], giving support to the percept [t], resulting in predominantly fusion T responses. In noise, the reliability of the acoustic signal decreased, and the visual weighting increased for both McGurk stimuli, resulting in more K responses.

To our knowledge, there is just one previous study in addition to the current one, which has investigated both acoustic and visual physical stimulus features contributing to the McGurk effect. In that study, the A and V features of fusion McGurk stimuli accounted for 52% of the variance in response distributions, when the visual features of consonants in [a] context were obtained by analysing the motion paths of optical reflectors placed on selected locations on the talker's face [15]. The relationship between the physical features and perception might be even stronger if a more elaborate measure of visual features were developed. Still, this finding emphasizes the importance of taking account the physical stimulus features when interpreting the McGurk effect, which is regrettably a largely ignored issue.

There are a couple of previous studies, which have concentrated on the role of some acoustic features in the McGurk effect. Green and coworkers [10-12] have conducted elegant studies showing that English-speaking listeners hear a fusion McGurk stimulus more frequently as [ð] in [a] context, and more frequently as [d] in [i] context. This is another example of a qualitative difference in the McGurk effect due to vowel identity. Green [11] suggested a similar explanation to ours involving F2 transitions, so that in [e] or [i] context the rising transition for bilabial and alveolar consonants makes them more confusable with each other than bilabials and velars. However, Green and coworkers did not analyse their acoustic stimuli to test this hypothesis. Recently, Hampson and coworkers have confirmed the finding of more fusion percepts in [i] than [a] context [13]. Furthermore, they analysed the F2 formants of the acoustic stimuli, providing similar results as our phonetic analysis, thus providing evidence for the role of F2 transitions in modulating the McGurk effect.

Release bursts have previously been shown to affect the McGurk effect so that an amplified burst of [b] results in more frequent fusion percepts [12]. As the spectrum was not analysed, the reason for this is unclear. Our phonetic analysis showed that the release burst of [ete] was wide-spectrum, which contributed to auditory [epe] being confused with [ete] in white noise.

In multisensory integration, the final percept depends on the reliability of the components [6]. The acoustic cues discriminating stop consonants are rather weak and sensitive to disturbance [16]. In the current study, when visual speech was relatively more reliable, it dominated the percept – particularly when the acoustic signal got poorer – so that the McGurk stimulus was mostly heard as [k] in [a] vowel context. When both visual and acoustic speech contained cues for [t], the fusion percept dominated so that the McGurk stimulus was mostly heard as [t] in [e] context. In both cases, the relative reliability of the acoustic and visual information determined the final percept, which thus varied across the SNR.

5. Conclusions

The McGurk effect was qualitatively different depending on the vowel context. A combination of an acoustic [p] and a visual [k] was heard predominantly as [k] in [a] context, but as [t] in [e] context. The visual influence became stronger with a decreasing acoustic signal-to-noise ratio. The acoustic and visual phonetic features contributed to the McGurk effect in proportion to their relative reliabilities.

6. Acknowledgements

This research was funded by the Academy of Finland (project numbers 12555673 and 1265610).

7. References

[1] Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. J Acoust Soc Am, 26(2), 212-215.

[2] McGurk & MacDonald (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

[3] Tiippana, K., Puharinen, H., Möttönen, R., & Sams, M. (2011). Sound location can influence audiovisual speech perception when spatial attention is manipulated. Seeing Perceiving, 24(1), 67-90.

[4] Tiippana, K., Hayes, E., Möttönen, R., Kraus, N., & Sams, M. (2010). The McGurk effect at various auditory signal-to-noise ratios in American and Finnish listeners. In Proceedings of the AVSP2010, 166-169.

[5] Tiippana, K., Andersen, T. A., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. Eur J Cogn Psychol, 16, 457-472.

[6] Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. Trends Cogn Sci, 8, 162-169.

[7] Boersma, P., and Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.42, retrieved 2 March 2013 from http://www.praat.org/

[8] Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. J Acoust Soc Am, 27, 769-773.

[9] Shigeno, S. (2000). Influence of vowel context on the audio-visual speech perception of voiced stop consonants. Jap Psychol Res, 42, 155-167.

[10] Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. Percept Psychophys, 50, 524-536.

[11] Green, K. P. (1996). The use of auditory and visual information in phonetic perception. In Stork D. G. and Hennecke M. E. (Eds): Speechreading by Humans and Machines: Models, Systems, and Applications. Springer, Berlin.

[12] Green, K. P. and Norrix, L. W. (1997). Acoustic clues to place of articulation and the McGurk effect: the role of release bursts, aspiration, and formant transitions. J Speech Lang Hear Res, 40, 646-665.

[13] Hampson, M., Guenther, F. H., Cohen, M. A., & Nieto-Castanon, A. (2003). Changes in the McGurk effect across phonetic contexts. Techn Rep Univ Boston.

[14] Benguerel, A.-P., & Pichora-Fuller, M. K. (1982). Coarticulation effects in lipreading. J Speech Hear Res, 25, 600-607.

[15] Jiang, J., & Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. J Exp Psychol Hum Percept Perform, 37(4), 1193-1209.

[16] Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. J Acoust Soc Am, 27(2), 338-352.