Seeing glee but hearing fear? Emotional McGurk effect in Swedish

Åsa Abelin

Department of Linguistics
University of Göteborg, Sweden
abelin@ling.gu.se

Abstract

Video and audio recordings of emotional expressions of a male speaker were used to construct conflicting stimuli in order to perform a McGurk experiment. The results are compared to an almost identical experiment with a female speaker, described in Abelin (2007) in order to see if the earlier results were reproduced. The results from both experiments show that in the McGurk condition the visual channel was in general more reliable than audio at conveying emotions, i. e. the results from Abelin (2007) were reproduced. Most frequently the listeners interpreted in accordance with the visual stimuli, or they heard an emotion which was not present in either video or audio. The emotions happiness and anger showed a dominance for visual perception, and the emotions fear, and to some extent surprise showed a preference for auditive perception, in the McGurk condition. No emotional dimension, i.e. the evaluative (positive–negative) was connected specifically to the visual or the auditive modality.

1. Introduction

Speech is seldom perceived only auditorily, but bimodally. This paper presents the results of a perception experiment of emotional Mc Gurk effect in Swedish and compares it to a similar experiment made by Abelin (2007). Studies on perception of emotional expressions have generally been made either in the visual or in the auditory domain, see Scherer (2003) and Rosenblum (2003) for overviews, and many studies have been done on multimodal stimuli, e.g. Abelin (2004), Beskow et al (2006), House (2007).

The McGurk effect has been widely studied for vowel and consonant perception since McGurk and McDonald (1976), and concerns the perception of conflicting visual and auditive input, see e.g. Massaro (1998a) or Traunmüller (2006).

The McGurk effect in emotional expressions has not been studied to the same extent as consonants and vowels, but there are some studies, with somewhat conflicting results, as concerns the perceptual dominance of the visual and auditory modalities and whether fusions occur (e.g. Fagel (2006), Massaro (1998b). Fagel (2006), who studied German, suggests that the evaluative meaning dimension (e.g. happy vs. angry) is perceived from the facial expression, while arousal (e. g. happy vs. content) is perceived in the voice. It is not necessarily so that emotions are discrete, but they were treated as such in this exeriment. For further discussion see Abelin (2007) and Abelin & Allwod (2002).

1.1. Research questions

The questions under study in this experiment are the following: 1. Is the auditory or the visual modality dominant in perception of emotions with conflicting auditory and visual stimuli? 2. Are different emotions or emotional dimensions, such as positive/negative connected to a certain modality?

2. Method

One Swedish male speaker was video and audio recorded using a MacBookPro built-in camera and microphone. He expressed the six basic emotions *happy*, *angry*, *surprised*, *afraid*, and *disgusted,* saying "hallo, hallo". These bimodal expressions of emotions were subjected to a perception test with two subjects. This showed that the recordings were successful.

The audio and the video for the six emotions were separated and then combined to form the 13 McGurk stimuli shown in Table 1. The first 11 of these are identical in combinations and in order with the stimuli of Abelin (2007).

The stimuli were presented to nine perceivers (aged 19–29 years). The test employed forced choice, in contrast to the experiment of Abelin (2007), which employed free choice. An additional test with forced choice for the stimuli of the earlier experiment of Abelin (2007) had showed no great differences in results from the free choice test.

The subjects were tested one by one in a quiet room. They attended to each stimulus 1–3 times and the experimenter controlled that they were both listening to and looking at each stimulus. They wrote a mark on an answering sheet where the could choose between the emotions *happy*, *angry*, *surprised*, *afraid*, *sad* and *disgusted*, but also had the opportunity of writing down other emotions if they felt the emotions of the answering sheet inappropriate. The "other" alternative makes statistical analysis more difficult, but was used in order to keep the results comparable to the earlier experiment with free choice.

Table 1: *The stimuli of the experiment.*

Stimulus nr	Video	Audio
1	happy	angry
2	surprised	afraid
3	angry	happy
4	afraid	surprised
5	angry	afraid
6	surprised	angry
7	happy	afraid
8	afraid	angry
9	surprised	happy
10	angry	surprised
11	afraid	happy
12	happy	disgusted
13	disgusted	surprised

2.1. Method of analysis

In the McGurk test the answers were analyzed as "video/visual" if the response was in accordance with the video stimulus, "audio/auditive" if the response was in accordance with the audio stimulus, and "other" if the response was not in accordance with intended emotion in either video or audio. The "other" alternative thus includes any of the other five emotions in the list, as well as listeners' own suggestions.

3. Results

The general result is that perceivers interpret either 1) in accordance with the visual stimuli, 2) in accordance with the auditive stimuli 3) as neither of the two, suggesting an interpretation of the stimulus as something else than what face or voice intended.

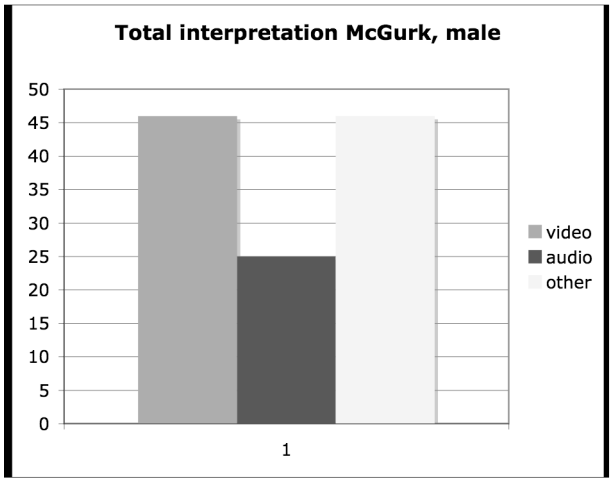

Figure 1. *The diagram shows listeners interpretations of the male speaker in accordance with video or audio stimuli, or as other.*

The diagram shows that perceivers generally interpret emotions in accordance with visual stimuli or as other feeling, more seldom in accordance with audio stimuli. Perceivers thus rely more on the face than on the voice in the McGurk condition. Also in the earlier experiment (Abelin, 2007), listeners least frequently interpreted stimuli in accordance with audio. However, in the earlier test "other" was the most common interpretation, see Figure 2.

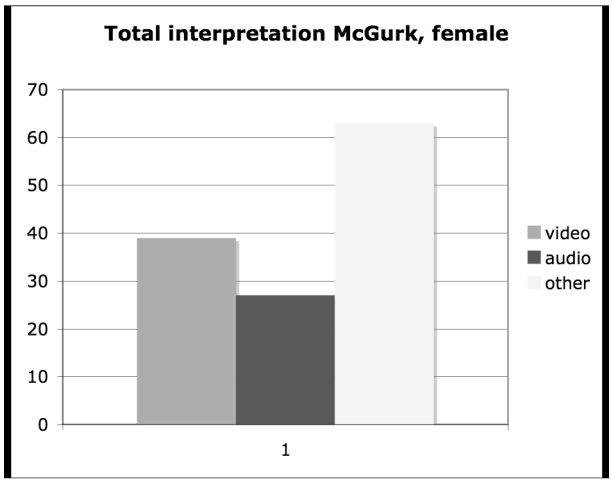

Figure 2. *The diagram shows listeners' interpretations of the female speaker in accordance with video or audio stimuli, or as other, in Abelin (2007).*

Figure 3 shows the interpretations of each of the 13 stimuli. The interpretations are mainly depending on vision for the visual stimuli happy, angry, surprised, disgusted and on audition for the auditive stimuli afraid and surprised. Surprised and angry thus has strong cues in both vision and audition while happy is more connected to vision and afraid more connected to audition.

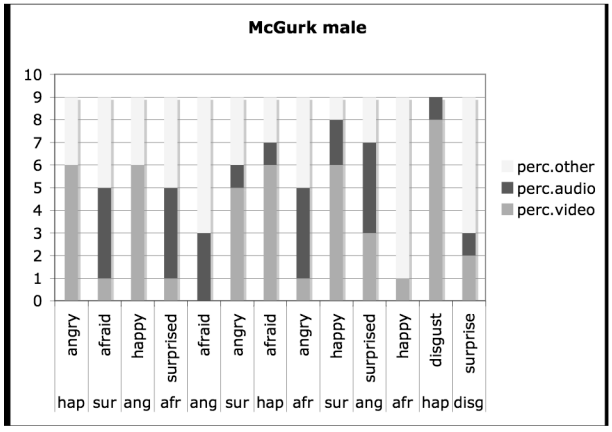

Figure 3. *The interpretations of each of the 13 conflicting stimuli. Only the first 11 are compared with the study of Abelin (2007).*

The second research question concerned whether any of the emotions or some emotional dimension is connected to a certain modality. Disregarding the "other" answers and only counting the number of responses which were in accordance with either the intended visual or auditory signal of a given stimulus we get the results shown in Figure 4. Except for the emotion afraid, the visual modality was interpreted more correctly than the auditive, generally. Furthermore, there was no division of modalities between positive (happy, surprised) or negative (angry, afraid) emotions.

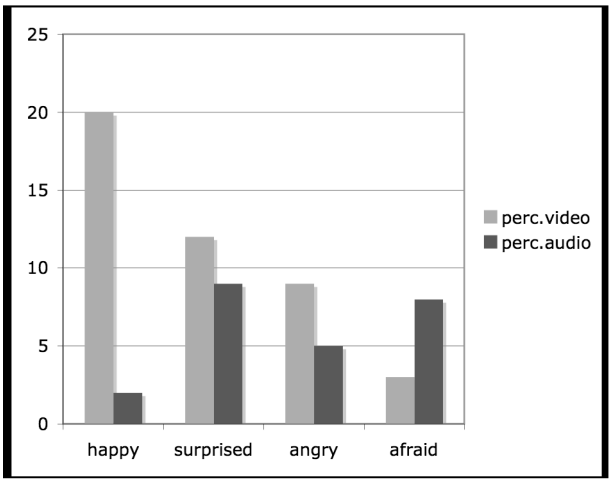

Figure 4. *The number of responses in accordance with the intended emotions in video and audio.*

The results shown in Figure 4 can be compared with the results of the earlier experiment, shown in Figure 5.

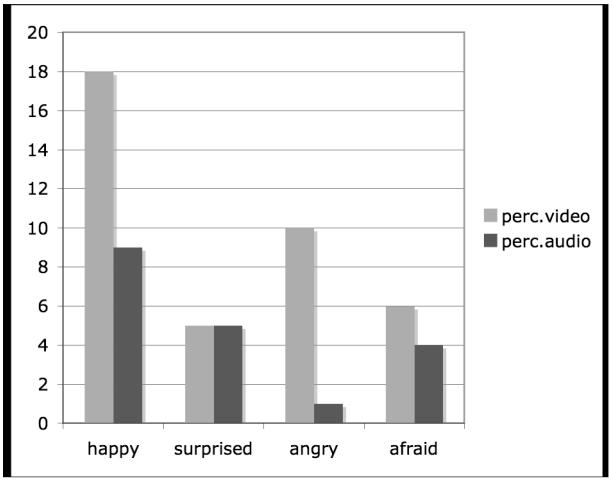

Figure 5. *The number of responses in accordance with intended stimuli, in video and audio, in Abelin (2007)*

The comparison shows a bias towards visually based interpretations for the emotions happy and angry in both studies. Surprise is more visually biased in the present study, while afraid is less visually biased and more auditively biased in the present study. The almost total dominance of the visual modality of the earlier study is weaker in the present study, since afraid is more auditive.

3.1. Summary of comparison with the earlier study

The main results from the McGurk test of Abelin (2007) were the following:

1) Perception in the McGurk condition was generally other emotions than the emotions displayed in the visual or the auditive channel. In second place came interpretation in accordance with the visual channel and, least common, interpretation in accordance with the auditive channel. It thus seems that in a situation with conflicting stimuli the visual channel is preferred.

2a) In general no emotions studied seemed to be connected to only a certain modality.

2b) There is no evident relation between sense modalities and meaning dimensions such as positive/negative, in the McGurk condition.

The results of 1), was partly replicated in the second experiment, cf. Figure 1 and 2. The most common interpretations were in accordance with the visual channel or as other, e. g. a fusion. The auditive channel carried the least weight in both experiments. The result of 2a) was partly replicated, the visual modality was the strongest for each emotion, except for afraid, where the auditive channel dominated, cf. Figure 4 and 5. The result of 2b) was replicated in the present experiment. The negative/positive dimension was not connected to either vision or audition.

On the level of each of the 11 individual stimuli, 7 of them showed similar results in the two tests, cf. Figure. 6 and Table 2.

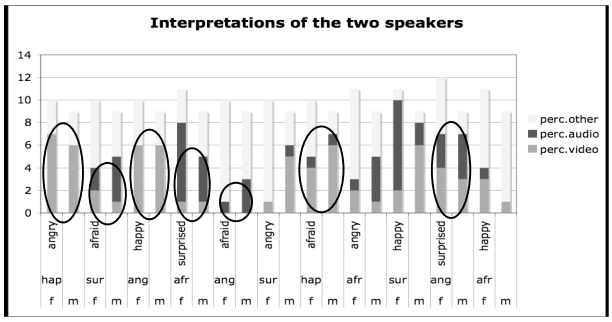

Figure 6. *Diagram of the results from the earlier and the present McGurk tests. Emotional labels in bottom row represent visual stimuli, emotional labels above (vertical text) represent auditive stimuli. First two bars represent the number of responses of the first stimulus; visual: happy and auditive: angry, for female and male speaker. Third and fourth bar represent number of responses of the second stimulus; visual: surprised and auditive: afraid, for female and male speaker, and so on.*

Figure 6 shows that seven of the stimuli: 1, 2, 3, 4, 5, 7 and 10 were similar in terms of proportions of video/audio/other interpretations for the two tests. These stimuli concerned either happy or angry as the correctly interpreted visual stimuli and surprised and afraid as the correctly interpreted auditive stimuli. Thus there are similarities for seven of the eleven stimuli. The proportions are shown in Table 2.

Table 2: *Percentual similarities for each emotion correctly responded to, in the two studies.*

Stimulus nr	Video	1st and 2nd study	Audio	1st and 2nd study
1	happy	70%, 67%	angry	
2	surprised	20%, 11%	afraid	20%, 44%
3	angry	60%, 67%	happy	
4	afraid	9%, 11%	surprised	64%, 44%
5	angry		afraid	10%, 33%
7	happy	40%, 67%	afraid	10%, 11%
10	angry	33%, 33%	surprised	25%, 44%

3.2. Differences in the interpretations of female and male speaker?

The similarities of the 7 stimuli in the two experiments, shown in Figure 6 and Table 2, can also be seen as similarities between interpretations of male and female speaker. However, this needs to be investigated further with more speakers of different sex. In both experiments the overwhelming majority of the listeners were young women. In the following experiments groups of young men will be used as listeners as well.

3.3. Conclusions and discussion

The conclusions from the two studies are the following:

The visual modality is generally dominant in perception of emotions with conflicting auditory and visual stimuli.

Some emotions seem to be more connected to vision, i.e. happy and angry, than are other emotions. Other emotions are interpreted better auditively, i.e. afraid, and especially in the first test, also surprised.

The tendency of dominance of different modalities for different emotions is stronger in the second study, but no emotion is completely dominated by the visual or to the auditory modality. Whether the stronger tendency in the second study is due to a difference in method or a difference in the sex of the speaker will be investigated further.

There is no evidence that emotional dimensions, such as positive/negative, are connected to a certain modality.

Neither of the modalities is connected to a certain emotion, which would of course be quite inefficient.

Thus there is also material for the discussion whether emotions are categorical or dimensional. Is there less evidence for emotional dimensions as a basic perceptual and psychological category since there seems to be no explicit modality connected to emotional dimensions? No, on the contrary they could be searched for in visual dimensions, such as eye and eyebrow movements, in the same way as emotional dimensions have been searched for in prosodic/acoustic dimensions, such as F0-variation, tempo and intensity (cf. Abelin & Allwood, 2002).

The title of this article touches upon the phenomenon of "hearing smiles" (cf. Aubergé, 2003). Clearly mouth movements are heard, as they are part of forming the acoustic signal, but the expression of happiness involves more than a smile; it involves the whole face (and body), maybe explaining why in some studies happiness is interpreted badly when only heard.

Why are some of the stimuli not interpreted in accordance with either the visual or the auditory component of the stinulus? In some cases fear is confused with sadness, or anger interpreted as irritation, emotions which are semantically similar (cf. Abelin & Allwood, 2002). In some cases blends occur, such as irritated from visual: happy–audio: disgusted, or tired from visual: surprised–audio: afraid. In other cases there can be contextual, individual, gender or methodological causes, which dimensions will be investigated further.

4. References

[1] Abelin, Å., 2004. Spanish and Swedish interpretations of Spanish and Swedish emotions – the influence of facial expressions. In *Proceedings of Fonetik 2004*. Stockholm, 108-111.

[2] Abelin, Å., 2007. Emotional McGurk effect in Swedish. In *Proceedings of Fonetik 2007, TMH-QPSR 50(1)*. Stockholm, 73-76.

[3] Abelin, Å.; Allwood, J., 2002. *Cross linguistic interpretation of emotional prosody*. Gothenburg papers in theoretical linguistics. Göteborg.

[4] Aubergé, V., 2003. Can we hear the prosody of a smile? *Speech Communication*, 40 (1-2), 87–97.

[5] Beskow, J.; Granström, B.; House, D. 2006. Focal accent and facial movements in expressive speech. *Proceedings from Fonetik 2006*. Lund, 9-12.

[6] Fagel, S., 2006. Emotional McGurk effect. *Proceedings from Speech Prosody 2006*, Dresden, 229-232.

[7] House, D., 2007. Integrating Audio and Visual Cues for Speaker Friendliness in Multimodal Speech Synthesis. *Proceedings of Interspeech 2007*. Antwerp, Belgium

[8] Massaro, D. W., 1998. *Perceiving talking faces: From speech perception to a behavioural principle.* Cambridge, Massachusetts: MIT Press.

[9] McGurk, H.; McDonald, I., 1976. Hearing lips and seeing voices. *Nature,* 264, 746-748.

[10] Rosenblum, L. D., 2005. Primacy of multimodal speech perception in: D B Pisoni and R E Remez, eds *The handbook of speech perception*, Blackwell publishing

[11] Scherer, K., 2003. Vocal communication of emotion: a review of research paradigms. *Speech Communication,* 40 (1), 227-256.

[12] Traunmüller, H., 2006. Cross-modal interactions in visual as opposed to auditory perception of vowels. In: G. Ambrazaitis, S. Schötz, eds, *Proceedings from Fonetik 2006*. Lund, 137-140.