Emotional McGurk effect in Swedish

Åsa Abelin
Department of Linguistics, Göteborg University

Abstract

Video and audio recordings of emotional expressions were used to construct conflicting stimuli in order to perform a McGurk experiment. Perception tests were made on bimodal, auditory, visual and McGurk stimuli. In unimodal condition the auditive channel showed more expected interpretations than the visual channel. In the McGurk condition however, the visual channel was more reliable than audio at conveying emotions. However, most frequently the listener hears something which is present neither in video nor audio. Bimodal non-conflicting information is most reliable. Meaning dimensions are relevant units for analysis, but neither emotional dimensions nor emotional concepts are connected specifically to the visual or the auditive modality.

Introduction

Speech is seldom perceived only auditorily, but bimodally. In fact, even when perceiving only the speech of a person we might be affected by earlier experiences of seeing that person articulate (Rosenblum 2005). This paper presents the results of a perception experiment of emotional Mc Gurk effect in Swedish. 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).

The phenomenon of McGurk has been widely studied 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. Fagel also found that the best unimodal results were obtained with audio only stimuli.

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? 2. Are the different emotions connected to a certain modality? 3. Are the modalities connected to certain meaning dimensions?

Method

One Swedish female speaker was video and audio recorded using a MacBookPro inbuilt camera and microphone. She expressed the six basic emotions *happy*, *angry*, *surprised*, *afraid*, *sad*, *disgusted* plus *neutral,* saying “hallo hallo”. These bimodal expressions of emotions were subjected to a perception test with five perceivers (Test 1). This showed that the most successful productions were *happy* (4 out of 5), *surprised* (5 out of 5) and *afraid* (4 out of 5). These three expressions were chosen, along with *angry* (which was generally interpreted as a negative emotion in the area *angry/irritated/serious/ sceptical*).

The audio and the video for these emotions were separated and then combined to form the McGurk stimuli shown in Table 1. There were no problems with the synchronization of audio and video. The stimuli were presented to ten perceivers (seven female, three male, aged 19–33 years) in test 2. In test 3 and test 4 audio and video were separately presented to a group of nine perceivers (all female, aged around 20 years). All tests employed free choice.

Method of analysis

As the subjects responded with free choice the answers had to be classified. As an example *afraid* included *insecure* and *worried* but not *sad*; *surprised* included *questioning* and *curious* but not *hesitating* and *angry* included *irritated* but not *serious*. In the McGurk test the answers were labeled "video" if the response was in accordance with the video stimulus, "audio" if the response was in accordance with the audio stimulus, and "other" if the response was not in accordance with either video or audio.

Results

Figure 1 shows the means for expected answers in three perceiving conditions; the results from test 1, 3 and 4. The bimodal condition yields the largest number of expected answers, the audio second best and the video condition comes in third place; however video produces more than 50% expected answers.

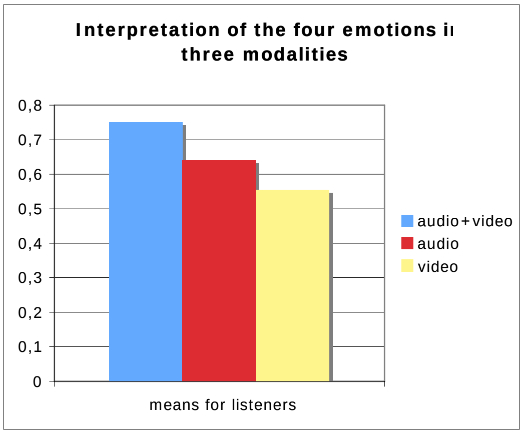

Figure 1. Interpretation of happiness, surprise, fear and anger in three conditions: bimodal, audio only and video only[1].

The stimuli of the McGurk test, test 2, were all the possible combinations of *happiness*, *fear*, *surprise* and *anger* and are shown in table 1.

Table 1. The stimuli of the McGurk condition.

Stim	Visual	Auditive
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	surprised

The general result from the McGurk test is that perceivers interpreted either 1) in accordance with the face, 2) in accordance with the voice 3) as other emotion which means a) a fusion: an interpretation of the stimulus as something else than what face or voice intended, or b) separation of voice and face which gave two different answers, however mostly not in accordance with either voice nor face.

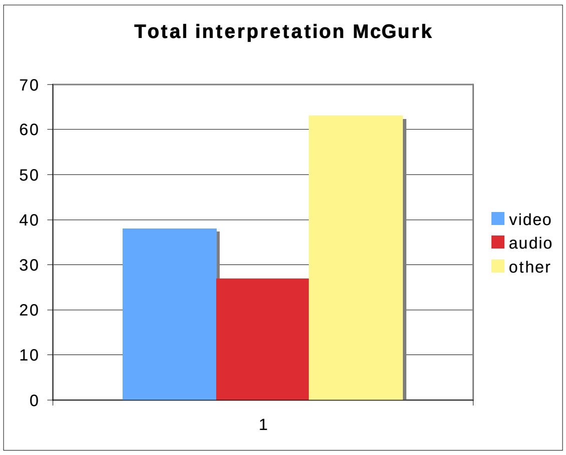

Figure 2. Perceivers generally interpret emotions as other than the video or audio, i.e. they do a fusion or a separation. Perceivers rely more on video than on audio in the McGurk condition.

The fusions and separations are mainly made when the emotions are not compatible, e.g. when *fear* and *anger* are combined (stimuli 5 and 8) or surprised face with angry voice (stimulus 6) see figure 3. When the emotions are compatible, e.g. *happy* and *surprise* (stimuli 9 and 12) , the interpretation is mainly done in accordance with the video or the audio channel. Compatibility can be characterized as having more meaning dimensions (see Abelin & Allwood, 2002) in common, e.g. +lust and +active for *happy* and *surprised*. Of the fusions, stimuli 5 (video: *angry* + audio: *afraid*) produced *neutral*, *hopelessness*, *insecure*, *sad*, *questioning*, *bad feeling*, *tired*, *molested*, *low*. All of them are –lust and –active.

[1] This was not due to the perceivers looking before listening, since there was a variation among the perceivers.

Figure 3 also shows that emotions such as *angry* and *happy* could be compatible.

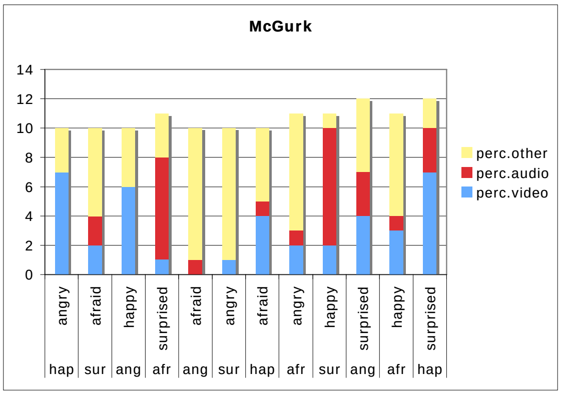

Figure 3. Interpretation of the twelve conflicting audio and video stimuli. Bottom row text: video, upper row text: audio.

Question number 2 was: Are the different emotions connected to a certain modality? Figure 4 shows that the emotions which were not perceived as “other” were generally perceived best visually. *Happy* and *angry* were the most visual emotions.

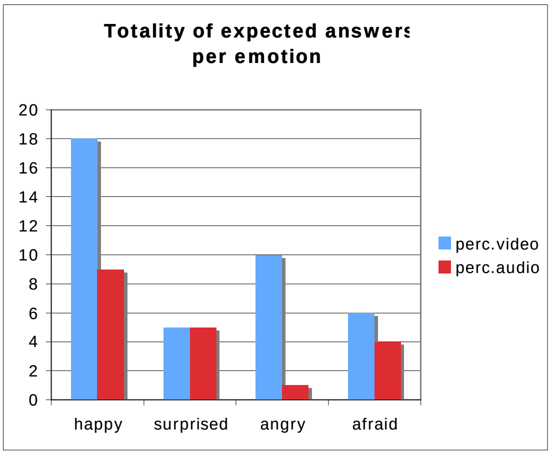

Figure 4. Interpretation of different emotions as expected, in video and audio, in the McGurk test. In a situation with conflicting stimuli the visual channel is generally preferred, but the dependency is different for different emotions.

When the stimuli were not interpreted in accordance with intended emotion, in video or audio, which emotions were interpreted better and worse? Figure 5 shows the number of “other” interpretations.

The audio channel was misinterpreted more often, except for *surprise*, but the differences were not so large. Figure 5 also shows that *happy* was generally better interpreted, independently of modality, while *angry* and *afraid* seem to be more difficult.

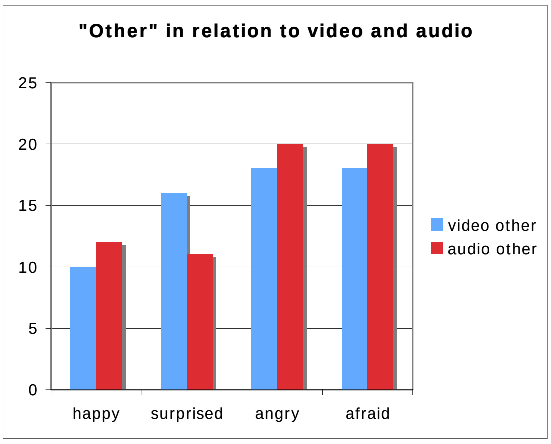

Figure 5. Analysis of interpretations “other” in relation to which emotion is presented in video and audio in the Mc Gurk test. The audio channel is somewhat more often misinterpreted than the video channel.

Concerning question 3, the emotions *happy*, *surprised* and *angry* are + active, but they were generally not interpreted better in the auditive modality (see Figure 6), as in Fagel´s study. In unimodal condition *happy* and *surprised* (both +lust) were not confused, neither in audio nor video except for one case in audio and one case in video (out of 36) where *surprised* was interpreted as *happy*.

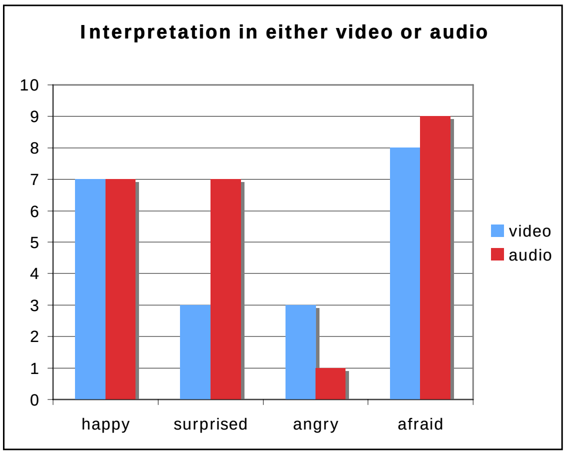

Figure 6. Interpretations of either video or audio. There is no evidence for meaning dimensions being connected to modalities.

Angry and *afraid* (–lust) were confused once, in audio, where *anger* was interpreted as *worried*. A positive dimension was never interpreted as a negative, or vice versa. Thus, there is evidence for meaning dimensions, but there is no relation between sense modalities and meaning dimensions. As concerns activity (afraid being –active and the other three +active) the results do not show the active ones being confused for each other. In the McGurk condition the emotions were confused more, both within and between dimensions. The emotions which were not perceived as “other” were perceived best visually. *Happy* and *angry* were most visual.

Analysis of meaning dimensions

Question 3 concerned whether the modalities were connected to certain meaning dimensions. The dimensions presented in Abelin and Allwood (2002) were +–lust, +–activity and +–secure. Using the concept of meaning dimensions one can understand how misinterpretations appear, as when *fear* gets interpreted as *sadness*; these emotions both have the meaning dimensions –lust, –activity. Likewise with *happiness* and *surprise*, which have + lust and +activity. There are of course more fine grained meaning components. There is no relation between sense modalities and meaning dimensions, neither in unimodal condition nor in McGurk condition.

Summary and discussion

The main results of the experiment were the following:

1a) Perception of the four emotions was most successful in normal bimodal condition, thereafter in auditive condition and least successful only visually (Figure 1).

1b) Perception in the Mc Gurk condition was generally a fusion, i. e. 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 (Figures 2 and 3).

2) In general no emotions studied seemed to be connected to a certain modality. On the contrary, happy was interpreted well independently of modality in the Mc Gurk test. In the auditive and the visual tests happy was also interpreted well (Figures 4 and 6).

3) There is no relation between sense modalities and meaning dimensions, neither in unimodal condition (Figure 6), nor in McGurk condition (Figure 4).

Emotions or emotional dimensions are perceived by hearing and vision. What is the relation between meaning dimensions and acoustic or visual dimensions? In Abelin & Allwood the attempt to match meaning dimensions with the acoustic variables F0, intensity and durations showed that there was far from a 100% one-to-one relation between semantic and acoustic variables, but some clear tendencies: + activity had greater F0 variation, stronger intensity and shorter durations, + lust had F0 variation (but that goes for anger as well). Instead of searching for just acoustic dimensions visual dimensions must be analyzed, as well as the context of communication.

In unimodal condition the auditive channel functions better than the visual in conveying emotions. In the mcGurk condition however, the visual channel works better. In case of conflicting information – invent something, otherwise trust the face! Bimodal non-conflicting information is the most reliable.

References

McGurk H McDonald I (1976). Hearing lips and seeing voices. *Nature,* 264, 746-748.

Abelin Å (2004). Spanish and Swedish Interpretations of Spanish and Swedish Emotions – the influence of facial expressions, in *Proceedings of Fonetik 2004*, Stockholm

Abelin Å Allwood J (2002). *Cross linguistic interpretation of emotional prosody.* Gothenburg papers in theoretical linguistics, Göteborg

Beskow J Granström B House D (2006). Focal accent and facial movements in expressive speech. In: G. Ambrazaitis, S. Schötz, eds, *Proceedings from Fonetik 2006*. Lund, 9-12.

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

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

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.

Fagel S (2006). Emotional McGurk effect. In: *Proceedings from Speech Prosody 2006,* Dresden.

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