Emotional McGurk Effect

Sascha Fagel

Institute for Speech and Communication
Technical University Berlin
sascha.fagel@tu-berlin.de

Abstract

Speaking is a physiological process that manifests in the acoustic and in the optic domain and hence is audible and visible. These two modalities influence each other in perception. Under normal circumstances the speech information in both channels is coherent and complementary and integrated to a percept. But if the information is conflicting and nevertheless integrated then the percept in one of the modalities might be changed by the other modality. The experiment described here discovers that when the video of an utterance spoken in one emotion is dubbed with the audio of the utterance spoken in another emotion the perceived emotion might be a third – neither present in the auditory nor in the visual modality.

1. Introduction

One of the most famous effects of sensory integration in speech perception – the so called McGurk effect – was described in 1976 by McGurk and MacDonald [7]. A syllable /da/ is frequently perceived when a video of the spoken syllable /ga/ is dubbed with the audio of the syllable /ba/. Since 1996 few researchers have investigated this effect in case of emotion perception in speech. Massaro and Egan [4] and Hietanen et al. [2] showed that when subjects are presented with an *angry* utterance in one modality a *happy* utterance in the other modality shifts the answers to *happy* and vice versa. De Gelder and Vroomen [1] achieved similar results with the combinations *happy-sad*, *angry-sad*, and *happy-afraid*. All these results were obtained with stimuli combined from two emotions or two emotions plus neutral in two (both emotions) or three alternatives (both emotions plus neutral) forced choice tests.

The experiment described in the following uses cross-combinations of four emotions in a four alternatives forced choice test. This approach enables the subjects – when presented with conflicting stimuli – to indicate a perceived emotion which is none of the emotions presented in the audio or in the video channel. The method of the present study was already applied by Massaro [5]: He tested audio-visually cross-combined stimuli of the word "please" (with natural audio and synthetic video) using the four emotions *happy*, *angry*, *surprised*, and *fearful*. A comparison of both studies is described in the discussion.

2. Method

2.1. Stimuli

An emotionally neutral sentence was chosen: "Dieses Teil bleibt übrig" ("This part remains"). In order not to deal with only one positive emotion these four emotions were employed in the experiment: two emotions with high arousal (*happy*, *angry*) and two emotions with low arousal (*content*, *sad*). Two of these emotions are positive (*content*, *happy*)

Figure 1: *Examples of the four emotions content, happy (top left and right), sad, and angry (bottom left and right). Note that in the whole video the dynamics transmits additional information and hence the visual stimuli are less ambiguous than still images.*

and two are negative (*sad*, *angry*). So these emotions span a wide area regarding the first two of the most commonly named emotion dimensions valence/pleasure and arousal/activation [8].

The sentence was uttered by a native speaker of German in the four emotions at approximately the same speech rate. The utterances were recorded in a sound proof booth with a Sony MiniDV camera with a high quality external clip microphone (video resolution 720·576 pixels at 25 frames per second and audio resolution 16 bit mono at 48 kHz). Figure 1 exemplarily shows frames of each of the four emotions.

The video track of each emotional utterance was dubbed with each audio track. Additionally the video and audio tracks were stored separately. This resulted in a total of 24 stimuli: 16 audiovisual (four coherent and twelve conflicting) and eight unimodal (four audio only and four visual only) stimuli. The stimuli can be examined at http://avspeech.info/EmoMcGurk.

2.2. Procedure

A pre-test revealed that if a subject is presented with an emotional speech video and has to rate it then if the same video track occurs a second time in the test the second rating depends on the first rating (the subject has already set its opinion). This is the case even if the video is combined with two different audio tracks: the rating is always dominated by the video that is presented first. The same applies for audio tracks. Therefore the test was split into six subtests whereas neither video nor audio track occurs twice within one subtest. Two of the subtests were composed of the eight unimodal stimuli (four audio only and four visual only) in two different orders. Each of the four audiovisual subtests contained three conflicting combinations and one coherent stimulus. In order to avoid adaptation and recalibration effects every stimulus was preceded by an audiovisual recording of the sentence uttered emotionally neutral. Each subtest was part of a public lecture of the author. The stimuli were presented one by one with a video beamer and loudspeakers in a lecture auditorium. Answers to the stimuli were collected on a paper form in a two alternatives forced choice procedure.

2.3. Subjects

The subjects were 412 visitors of an open day at the Technical University Berlin. Only answer sheets with all stimuli uniquely rated were analyzed. Data of 387 subjects (51% females, 49% males; age from 9 to 71 years, mean 37.5 years) remained.

3. Results

3.1. Unimodal stimuli

The audio only stimuli were less ambiguous than the visual only stimuli. Table 1 shows the confusion matrices. Except *angry* in visual only condition which was more often perceived as *sad*, and *content* in the visual only condition which was often confused with *happy* all intended emotions were identified at least at 80%.

		perceived emotion			
	intended emotion	*content*	*happy*	*sad*	*angry*
audio only	*content*	**86**	8	2	4
	happy	7	**93**	0	0
	sad	0	3	**96**	1
	angry	0	0	8	**92**
visual only	*content*	**59**	39	2	0
	happy	6	**93**	0	1
	sad	5	0	**80**	15
	angry	14	0	54	**32**

Table 1: *Confusion matrices for the conditions audio only (top) and visual only (bottom). Bold figures denote the identification of the intended emotion.*

3.2. Coherent audiovisual stimuli

Except for the emotion *content* whose identification score is between that of vision and audition the intended emotions of all coherent audiovisual stimuli are more often identified than in both unimodal conditions. Table 2 shows the confusion matrix for the coherent audiovisual stimuli. The audiovisual identification of the emotion *content* is roughly the same as in the visual only condition and only slightly enhanced by audition. The audiovisual identification of the emotion *angry* is not "disturbed" by the confusion with *sad* that occurs in the visual only condition.

3.3. Conflicting stimuli

As could be expected the results for conflicting stimuli are less clear than those for coherent audiovisual or unimodal stimuli. Table 3 shows the confusion matrix where the highest score for every stimulus is displayed in bold face. Three of the five values above 70% (98%, 85%, and 72%) show responses equal to the intended emotion in the audio part of the stimuli (marked with

	intended emotion	perceived emotion *content*	*happy*	*sad*	*angry*
audiovisual	*content*	**61**	38	0	1
	happy	6	**94**	0	0
	sad	0	0	**100**	0
	angry	0	0	4	**96**

Table 2: *Confusion matrix for the coherent audiovisual stimuli. Bold figures denote the identification of the intended emotion.*

(*)). For the other two high values the answer is neither the intended emotion in the audio nor in the video part of the stimuli (marked with (**)): 78% of the subjects perceived a *sad* utterance when *content* audio was combined with *angry* video and 74% of the subjects perceived a *happy* utterance when *angry* audio was combined with *content* video. The most frequent answers to the seven remaining stimuli are below 60% and show responses equal to the video part in five cases, equal to the audio part in two cases.

intended emotion audio	visual	perceived emotion *content*	*happy*	*sad*	*angry*
content	*happy*	38	**56**	4	2
	sad	39	0	**48**	13
	angry	3	0	**78** (**)	19
happy	*content*	15	**85** (*)	0	0
	sad	20	**51**	27	2
	angry	20	**45**	20	15
sad	*content*	**49**	25	24	2
	happy	35	**52**	13	0
	angry	2	0	**98** (*)	0
angry	*content*	23	**74** (**)	0	3
	happy	40	**42**	3	15
	sad	4	0	24	**72** (*)

Table 3: *Confusion matrix for the conflicting audiovisual stimuli. The highest score for every stimulus is displayed in bold face. Scores above 70% are marked with (*) or (**), respectively (see text for details).*

The distribution of answers to conflicting audiovisual stimuli can further be viewed in relation to the identification rate of the both stimulus parts in unimodal condition. For seven of the twelve conflicting stimuli that emotion is preferred which is more often identified (i.e. less ambiguous) in the unimodal condition. For five stimuli this is not the case. This indicates that bimodal emotion perception does not mean simply selecting the most informative channel.

3.4. Robustness of the modalities

The top of figure 2 shows the identification scores for the auditory modality when no video is displayed and when a conflicting video (with a different emotion than in the audio part of the stimulus) is played with the audio. The same is shown for the visual modality (bottom). Both modalities are about equally resistant against conflicting information in the other modality which indicates about equal influences from the two modalities. But for the auditory modality *content* and *angry* (which are least frequently identified in unimodal condition) are more susceptible to distortion by conflicting information than *happy* and *sad* whereas for the visual modality *content* and *angry* (which are also least frequently identified in unimodal condition) are less susceptible than *happy* and *sad*.

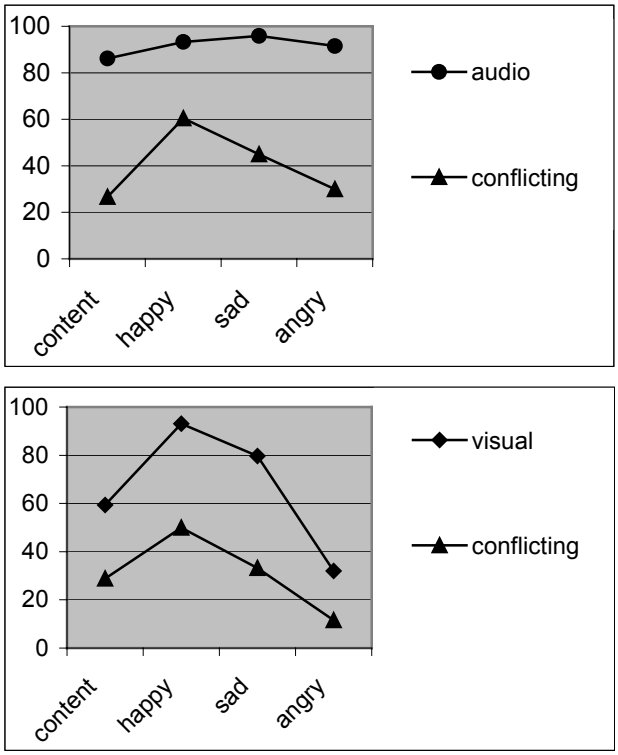

Figure 2: *Percent correctly identified stimuli as a function of the intended emotion. Top: audio and conflicting (with a different emotion presented in the visual modality) audiovisual stimuli. Bottom: visual and conflicting (with a different emotion presented in the auditory modality) audiovisual stimuli.*

4. Discussion

In Massaro's study [5] of audio-visually cross-combined stimuli (the word "please" in four emotions with natural audio and synthetic video) only one barely distinct answer that did not reflect the audio or visual part of the stimulus was given most frequently to a stimulus: ca.

41% of his subjects reported to perceive *surprised* when *happy* audio was played along with a *fearful* video. Answers to *surprised* audio with *fearful* video reflected the audio part as well as all stimuli with *fearful* audio. Answers to all other stimuli showed the video answer. Furthermore, except for the aforementioned *happy* audio with *fearful* video stimulus, for all other conflicting stimuli that answer was given most frequently that reflects the modality which was less ambiguous in the unimodal condition. Hence, integration of the auditory and visual sources of information regarding emotion did hardly occur. This might be due to the fact that Massaro presented the same audio and video parts more than once in a trial to the subjects. Another explanation might be the discrepancy between the audio (natural) and the video (synthetic) stimulus parts.

MacDonald and McGurk [3] early stated a manner/place theory of human speech recognition which said that the manner of articulation is transmitted auditorily and the place of articulation is transmitted visually. Although this strong formulation is nowadays defeated [6] the information of the manner of articulation is hardly present in the video channel and the information of the place of articulation in the audio channel might easily be disturbed by noise. The results of the present experiment show a similar tendency in the case of emotion perception. Assuming that the video channel primarily transmits the valence (if the face shows a positive or a negative emotion) and the audio channel primarily reflects the arousal (if the speaker is more or less excited) although both kinds of information are to a certain extend present in both channels this would lead to several conclusions: a positive (e.g. *content*) face with an exited (e.g. *angry*) voice leads to the perception of a *happy* utterance. This scheme applies for eight of the conflicting stimuli. For three of the remaining that part of the stimulus that was more clearly identified in the unimodal condition dominates the perception. For one stimulus this scheme fails. But for this stimulus (*sad* audio with *happy* video) nonetheless a remarkably high number of fusion answers (*content*) according to the scheme occurs.

The results of the present study cannot be explained by simply selecting the less ambiguous source of information even for answers that reflect the intended emotion of one stimulus part in unimodal condition. In fact answers are frequently given that represent a third emotion which is not present in one of the modalities. A model of perception should be applied to the data obtained by the experiment. But the commonly used fuzzy logical model of perception (FLMP, [5]) has been claimed not to be applicable to data derived from conflicting stimuli [9]. The extended model – the weighted FLMP (WFLMP, [10]) – does not work here due to the necessary splitting into subtests. The perception of emotion includes many cues that also might depend on each other. Not only the dimensions of emotions regarded in the present experiment, namely valence and arousal, but other dimensions might be candidates for these cues that contribute to the perception of emotion. Experiments with more emotions than *content*, *happy*, *sad*, and *angry* shall be carried out. Low level cues like e.g. voice quality parameters and articulation features may also be taken into account. The present experiment suggests that these cues are weighted in the modalities audition and vision and integrated to the perception of an emotion that does not necessarily exist in the presentation of only one of the stimulus' modalities.

5. References

[1] de Gelder, B., Vroomen, J., 2000. The perception of emotions by ear and by eye. *Cognition & Emotion* 14(3), 289-311.

[2] Hietanen, J.K., Leppänen, J.M., Illi, M., Surakka, V., 2004. Evidence for the integration of audiovisual emotional information at the perceptual level of processing. *European Journal of Cognitive Psychology* 16, 769-790.

[3] MacDonald, I., McGurk, H., 1978. Visual Influences on Speech Perception Process. *Perception & Psychophysics* 24, 253-257.

[4] Massaro, D.W., Egan, P.B., 1996. Perceiving Affect from the Voice and the Face. *Psychonomic Bulletin and Review* 3(2), 215-221.

[5] Massaro, D.W., 1998a. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. Cambridge, Massachusetts: MIT Press.

[6] Massaro, D. W. 1998b. Illusions and Issues in Bimodal Speech Perception. *Proccedings of the International Conference on Audio-Visual Speech Processing*, Sydney, 21-26.

[7] McGurk, H., MacDonald, I., 1976. Hearing Lips and Seeing Voices. *Nature* 264, 746-748.

[8] Russell, J., Feldman Barrett, L., 1999. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology* 76, 805-819.

[9] Schwartz, J.-L., 2003. Why the FLMP should not be applied to McGurk data ... Or how to better compare models in the Bayesian framework. *Proceedings of the International Conference on Audio-Visual Speech Processing*, St. Jorioz, 77-82.

[10] Schwartz, J.-L., Cathiard, M., 2004. Modeling Audio-Visual Speech Perception: Back on Fusion Architectures and Fusion Control. *Proceedings of the 8th International Conference on Spoken Language Processing*, Jeju, 2017-2020.