Advances in Life Science and Technology
ISSN 2224-7181 (Paper) ISSN 2225-062X (Online)
Vol 4, 2012

An Auditory-Visual Conflict of Emotions - Evidence from McGurk Effect

Sudhin Karuppali*, Jayashree S. Bhat, Krupa P, Shreya

Department of Audiology and Speech Language Pathology, Kasturba Medical College (A Unit of Manipal University), Mangalore

* E-mail of the corresponding author: sudhin.karuppali@manipal.edu

Abstract

Experiments such as the McGurk effect have supported the inter-dependence of the auditory and visual modalities. This perceptive reasoning is in line with the assumption that emotional expressions in the face and voice are processed by the same perceptual/cognitive mechanism. This research aimed to study the accuracy of the identification of intended emotions in the Kannada language using the auditory and/or visual modality; and also to study its perception using the McGurk paradigm. An emotionally neutral word "namaskara" was uttered by a native Kannada speaker in four basic emotions. Subjects were asked to identify the intended emotional expressions under the different experimental conditions (unimodal, bimodal-I, bimodal-II). The scores were recorded and analysed respectively. The results were in line with studies who also stated the over-reliance of visual over auditory modality when the subjects were presented with the McGurk stimuli thereby also perceiving a new different emotion.

Key words: McGurk effect, bimodal, emotional, Kannada, stimuli

1. Introduction

Visual as well as auditory information, if available are imperative for speech perception. Seeing the speaker's face can significantly improve the intelligibility of speech especially if the speech has to occur in a noisy environment (Calvert et al, 1997; Macaluso et al, 2004). The importance of facial movements for speech perception has gained wide recognition with the work of McGurk and MacDonald (1976), showing that speech information from the face that is incompatible, leads to misleading percepts. These visual cues have a broad influence on the perceived auditory stimuli. The McGurk effect, as the phenomenon is generally called, adds to a body of knowledge which includes cross-modal interactions observed in the localization of sounds (Bertelson, 1998), suggesting that speech perception is multimodal. Here the visual articulatory information is integrated into our perception of speech automatically and effortlessly. The syllable that we perceive depends on the strength of the auditory and visual information. This visual information is thought to influence the perception of speech because associations between visual features and phonological representations have been acquired through the experience of watching the talker's mouths move while listening to them speak. According to Summerfield (1987), having both auditory and visual information assists in the discrimination between consonants that are easily confused when presented in just one modality, which is also observed for all languages (Massaro, Cohen, Gesi, Heredia, and Tsuzaki, 1993).

Likewise in everyday life, the perception of emotions appears to also be bimodal. According to Frijda (1989), Darwin was one among the first theoreticians of emotion to consider emotions as closely related to action. Darwin (1872) and Ekman (1982) postulated the predominance of facial signs in human perception of emotion and its relative advantages over modalities like speech. Perception of either an angry face or an angry voice generally leads to the conclusion that the person is angry, not that his face looks angry, nor for that matter that his voice sounds angry. This perceptive reasoning is in line with the assumption that emotions in the face as well as emotional expressions in voice are processed by the same perceptual or cognitive mechanism. A study was carried out by De Gelder and Vroomen (2000), in which

they asked their subjects to identify an emotion, given a photograph and/or an auditory spoken sentence. They found that identification judgments were influenced by both sources of information. Also, Massaro (2000) who considers speech and emotion perception require multiple sources of information, made experiments with an animated talking head expressing four emotions in both unimodal and bimodal conditions (auditory & visual modalities). He found that his subjects attained an overall accurate performance when both the sources of information were present. Abelin (2004) conducted a cross linguistic experiment on multimodal interpretation of emotional expressions. He presented audio recordings of Spanish emotional expressions to Spanish and Swedish listeners. The results indicate that intra-linguistic as well as cross-linguistic interpretation of emotional prosody was improved by visual stimuli and seemed to be greatly augmented by multimodal stimuli.

Understanding and interpreting emotional expressions have always been important for effective communication to take place. It would be interesting to know the extent of auditory-visual interaction in an Indian language. Hence, the present study explores the effects of a perceptual experiment of emotional McGurk effect in Kannada (a south Indian language). The aim of the study was to study the accuracy of identification of the intended emotions using the auditory and/or visual modality; and also to study the auditory – visual conflict in the perception of different emotions in Kannada language, by using the conflicting stimuli constructed for the McGurk experiment.

2. Method

The present study was conducted at the Kasturba Medical College, Mangalore. Institutional ethical board approved the study and the informed consent was obtained from all the subjects before the commencement.

2.1. Stimuli preparation

An emotionally neutral word ‘namaskara’ (which means ‘hello’ in the Kannada language) was chosen as the stimulus. Four basic emotions were employed in the experiment. The emotions ‘happiness’ and ‘surprise’ were considered to be positive and ‘sadness’ and ‘anger’ as negative emotions. A native Kannada speaker uttered the chosen word in these four emotions maintaining the same speech rate. These utterances were recorded in a sound treated room with a Sony Handicam DCR-DVD610E (video resolution 0.34 megapixels and audio resolution 16 bit mono at 48 kHz). Each video track of an utterance was dubbed with each of the audio track respectively. In addition to it, both these tracks were isolated and stored separately. Thus, a total of 24 stimuli were obtained. The prepared stimuli were used in three experimental conditions namely – unimodal (audio only and video only stimuli), bimodal-I (coherent stimuli) and bimodal-II (conflicting stimuli). Coherent stimuli refers to the authentic emotional expression with both the audio and video information; whereas the conflicting stimuli or the McGurk stimuli consists of combinations of audio and video information (extracted for the unimodal condition) of the different emotional expressions. The extracted stimuli were presented in two experimental methods – Test 1 and Test 2 (Table 1).

2.2. Subjects

Test 1 consisted of 10 subjects (5 males and 5 females) and Test 2 consisted of 18 subjects (9 males and 9 females). All of the subjects were native Kannada speakers within an age range of 18-25 years. None of the subjects had any history of auditory or visual deficits and had a minimum qualification of secondary education.

2.3. Procedure

Test 1

Ten native Kannada speakers judged the recorded stimuli from two conditions - unimodal and bimodal-I.

Each stimulus was presented via a standard compatible laptop (Lenovo Y410) and a loudspeaker connected to its port. The subjects were instructed to identify the intended emotion in each of the stimuli. They were given a closed set of stimulus choices (to prevent varied descriptions for each emotion), with only the four target emotions.

Test 2

Eighteen native Kannada speakers were asked to identify the intended emotions from the bimodal-II condition. The subjects were free to give an open ended response. Any varied descriptions were also considered as a correct response, as long as their emotional perception encircled around being either a positive or a negative emotion as mentioned earlier.

2.4. Method of analysis

Responses to the stimuli were collected on a paper. In Test 1, the percentage of accurate identification of the intended emotion under the two unimodal and bimodal-I conditions were recorded. And for Test 2, the percentage of interpretation of the stimuli were recorded for the bimodal-II condition, which could either be in accordance to the face, voice or any new emotion other than what face or voice intended to. The responses were labeled “face” if the response was in accordance with the visual stimulus; “voice” if the response was in accordance with the auditory stimulus and “other” if the response was not in accordance with either visual or auditory, resulting in a new perceived emotion. The results were analyzed using SPSS (11.5). A descriptive statistics was employed for the purpose.

3. Results

The interdependence of auditory and visual information for effective speech perception is an established fact. However, this dependency with respect to emotional expressions is a less explored area. A visual emotion interposed with a different auditory emotion makes the listener perceive the output stimuli to be a totally different emotion, further establishing the notion of having a common cognitive-perceptual processing system. The present study focuses on such an aspect. An emotional expression was recorded in four different states (happiness, surprise, sadness and anger) and in four conditions (unimodal-audio, unimodal-video, bimodal-I and bimodal-II) and was judged by native Kannada speakers. The analysis made was based upon the percentage of accurate identification of the intended emotion in the unimodal and bimodal-I conditions and the percentage of reliance on either visual or auditory information in the bimodal-II condition (Table 2).

Test 1

For the unimodal (audio only) condition, the emotional expressions were perceived to be the intended emotion 90% of the time on an average and for the unimodal (video only) condition it was less than 85% of the time. The overall accuracy of identification of the intended emotion was better for the auditory mode when compared to the visual mode. However in the bimodal-I condition, the stimuli were identified much more accurately than either of the unimodal conditions with more than 94% accuracy. Table 2 shows the percentage of accuracy of identifying the emotional expressions in the different conditions.

Test 2

The results of the bimodal-II condition were entirely different. Unlike the unimodal conditions in which better responses were obtained in the audio mode than the video mode, the bimodal-II condition revealed a reverse effect. The visual information was more relied upon than the auditory information in the

perception of the McGurk stimuli. The results demonstrated that subjects interpreted either 1) In accordance with the face (63.52%); 2) In accordance with the voice (16.37%); 3) or as a new emotion (20.11%). The emotions which were not perceived as "other" was perceived best visually. Though the dependency on visual cue is appreciably higher than the auditory cue, the percentage of dependency is different for different emotions. The emotions 'surprise' and 'anger' were identified best by visual mode. Figure 1 shows a graphical representation of the percentage of dependency of 'voice' and 'face' for the identification of the emotional expressions.

4. Discussion

The present study was carried out to determine the degree of reliance of either the auditory or visual modality during the perception of an emotional utterance. In test 1, a higher accuracy in the perception of an intended emotion was achieved in the unimodal (audio) than the unimodal (video) condition, revealing an over-reliance on the auditory information than the visual information. The bimodal-I condition revealed an enhancement in the emotional perception abilities. This may be due to the close and intricate interaction of both the visual and auditory modalities. In accordance with these findings, De Gelder & Vroomen (2000) have commented that the identification judgements of the intended emotion depended on both the auditory and visual sources of information. Eventually Massaro (2000) also reported an overall accurate performance in bimodal conditions. The percentage of higher accuracy obtained for happiness and sadness in our study could be because that these two are the most contrasting emotions which could easily be differentiated.

In contrast to bimodal-I, the bimodal-II condition revealed a higher reliance on auditory than visual information. Mehrabian & Ferris (1967) found that the facial information was three by two times more important than the vocal information. Their results posit that the face is more important than voice for judging a portrait emotion (Bugenthal, Kaswan, Love, & Fox, 1970; Hess, Kappas, & Scherer, 1988). In our study, the percentage of reliance on visual information in the bimodal-II condition was almost 5 times more than that of the auditory information, especially in perceiving the emotion of 'anger' and 'surprise'. Fagel (2006) also suggested that the confusion in identifying the intended emotion was observed to be more in video only than in audio only condition, better in the bimodal condition and poorer in the McGurk condition. Massaro & Egan (1996) presented the subjects with a single word recorded by a speaker in a conflicting condition and found that the frequency of the response depended on the emotions expressed in both the face and the voice. However in the McGurk condition, they found that the visual channel was more reliable than auditory at conveying emotions. In our study, anger and surprise attained atleast 80% of dependency on the visual information when compared to 10% for the auditory information. Happiness and sadness also attained a higher percentage of dependency of visual over auditory information.

In the current study the perception of different emotions were observed 20% of the time. Abelin (2007) found mismatches between the visual and auditory displays of emotion that resulted in an `emotional McGurk' effect thereby creating the perception of a third, different emotion, which was also supported by Hietanen et al (2004) who attained similar results. As discussed earlier, the findings by Abelin (2004) also demonstrated the reliance of visual information; however, his study was in contrast with the results of the present study, in which the 'anger' and 'surprise' emotions were the most interpreted emotions on visual mode attaining more than 70% accurate identification scores, as compared to the other two emotions. Many researchers opine 'anger' as the most visual emotion (Abelin and Allwood, 2000; Scherer, Banse, and Wallbott, 2001; Abelin, 2007). Matsumoto et al (2002) reviewed studies on the influence of culture on the perception of emotion, and concluded that there is universality as well as culture-specificity in the perception of emotion. This research has shown that speech perception is greatly influenced by visual perception. Assuming that there is interaction between the senses, and that facial expression of emotion is more universal than prosody is, then cross-linguistic interpretation of emotions should be more successful

multimodally than only vocally.

5. Conclusion

Auditory-visual integration in speech perception has been a well-researched area right from the time of McGurk and MacDonald in the 1970s. The implication of their results was also superimposed in the perception of emotional expressions which is relatively a new area of research. The present study focuses upon the importance of intricate interaction between the auditory and visual modalities for the perception of emotional expressions. Results revealed the over-reliance of visual over auditory modality when the subjects were presented with the McGurk stimuli thereby also perceiving a new different emotion. But this dependency was observed to be vice-versa for the perception of the actual stimulus. Having a larger sample size could have increased the validity of the study. The present study can also be applied to other Indian languages and also to study the organization of the cognitive-perceptual processing system in bilinguals and multilinguals. Whether or not voice and facial expression recognition are organised around primitives or basic emotions, and whether or not these are the same in two cases are questions for future research. In this present state of our knowledge, various options must remain open. There are more events that signal potentially relevant information in our environment than just the movements of human face. Is audiovisual emotion perception really more like another case of the McGurk illusion, operating over inputs provided by the sights and the sounds of the face is still a debatable issue.

6. Acknowledgment

We are grateful and thank our subjects for their cooperation to conduct this study and making it a good success. We thank the Dean, Kasturba Medical College, Mangalore (A Unit of Manipal University) for supporting us to conduct such a study.

References

Abelin, A. (2004). Spanish and Swedish interpretations of Spanish and Swedish Emotions - The influence of facial expressions, in *Proceedings of Fonetik, 2004*, Stockholm.

Abelin, A. (2007). Emotional McGurk effect - An experiment. In *7th International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*.

Abelin, A., & Allwood, J. (2000). Cross linguistic interpretation of emotional prosody. *ISCA Workshop on Speech and Emotion. Newcastle*, Northern Ireland, 110–113.

Bertelson, P. (1998). Starting from the ventriloquism: The perception of multimodal events. In Sabourin, M., Craik, F.I.M., & Roberts. M. (Eds), *Advances in psychological science: Vol. 1.Biological and cognitive aspects*. Hove, UK: Psychology Press.

Bugenthal, D.E., Kaswan, J.W., Love, L.R., & Fox, M.N. (1970). Child versus adult perception of evaluative messages in verbal, vocal, and visual channels.*Developmental Psychology, 2*, 367-375.

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., & David, A.S. (1997) Activation of auditory cortex during silent lip read. *Science, 276*, 593-596.

Darwin, C. (1872). *The Expression of The Emotions in Man and Animals*, London: Murray

De Gelder, B. &Vroomen, J., (2000). The perception of emotions by ear and eye. *Cognition and Emotion*, 14, 289–311.

Ekman, P. (1982). *Emotion in the Human Face*. Cambridge Univ. Press, Cambridge

Fagel, S (2006). Emotional McGurk Effect. *Proceedings of the Speech Prosody conference*, Dresden.

Frijda, N. (1989). *The emotions.* Cambridge: CUP.

Hess, U., Kappas, A., & Scherer, K. (1988). Multichannel communication of emotion: synthetic signal production. In Scherer, K. (Ed), *Facets of emotion: Recent research*, 161-182.

Hietanen, J.K., Leppanen, 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.

Macaluso, E., George, N., Dolan, R., & Spence, C. (2004). Spatial and Temporal factors during processing of auditory visual speech- a PET study. *Neuroimage, 21*, 725-732.

Massaro, D. W., (2000). Multimodal emotion perception: Analogous to speech processes. *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, Northern Ireland, 114–121.

Massaro, D.W., Cohen, M.M., Gesi, A., Heredia, R., &Tsuzaki, M. (1993). Bimodal Speech Perception: An Examination across Languages. *Journal of Phonetics, 21*, 445-478.

Massaro, D.W., & Egan, P.B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin and Review, 3*, 215-221.

Matsumoto, D., Franklin, B., Choi, J.-W., Rogers, D. Tatani, H., (2002). Cultural influences on the Expression and Perception of Emotion" in W.B. Gudykunst and B. Moody, Eds. *Handbook of International and Intercultural Communication*, Sage Publications.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

Mehrabian, A., & Ferris, S.R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology, 31*, 248-252.

Scherer, K. R., Banse, R., and Wallbott, H. G., 2001. Emotion inferences from vocal expression correlate across languages and cultures, *Journal of Cross-Cultural Psychology*, 32 (1), 76–92

Tables

Experimental Methods	Experimental conditions	Stimuli	Number of stimuli	Emotional Expressions
Test 1	Unimodal	Auditory	4	Happiness
		Visual	4	Surprise
	Bimodal - I	Coherent	4	Sadness
Test 2	Bimodal - II	Conflicting	12	Anger

Table 1: The types and number of stimuli prepared for different experimental conditions.

	Unimodal conditions		Bimodal-I condition
	Audio only	Visual only	Coherent stimuli
Surprise	92%	77%	97%
Happiness	95%	83%	98%
Sadness	89%	81%	100%
Angry	91%	67%	94%

Table 2: Percentage of accuracy for the identification of the emotional expressions in the unimodal and bimodal-I conditions.

Figures

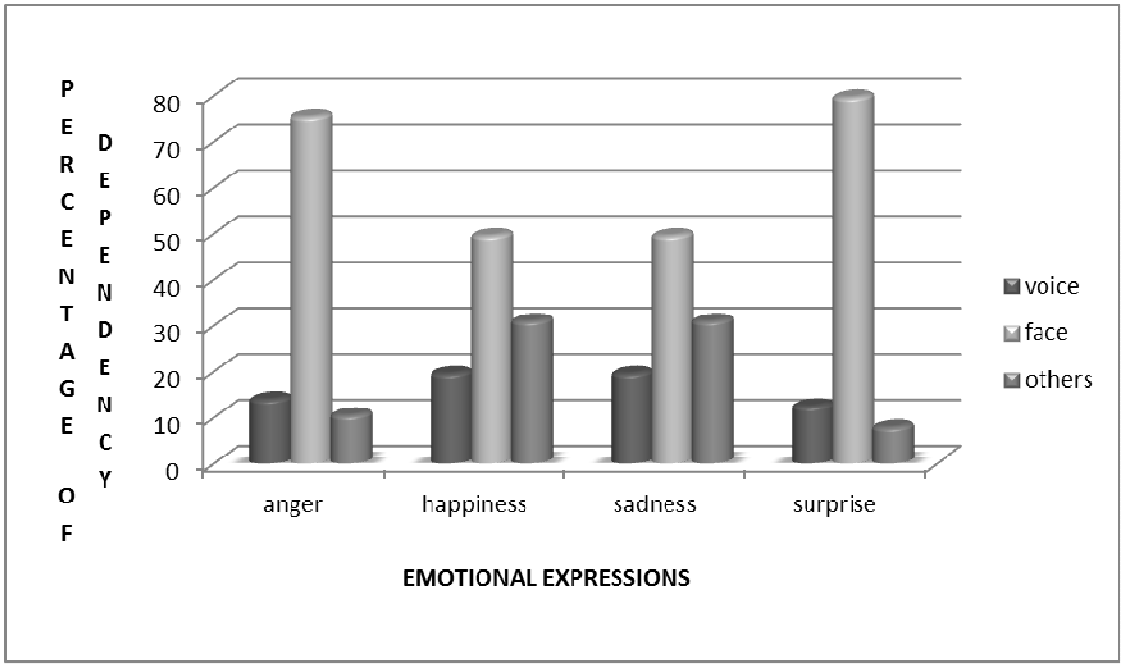

Figure 1: Percentage of dependence on visual cues, auditory cues or perception of a new emotion in the bimodal-II condition