Speech Perception and Emotion: Linguistic vs. Emotional Content

Abigail Montes De Oca , Theresa Cook, James Dias, Larry Rosenblum

Department of Psychology
University of California, Riverside

ABSTRACT

Individuals perceive speech through at least two sensory modalities: audition (hearing) and vision (seeing). We wanted to determine whether people perceptually prioritize emotional information or linguistic information when they see and hear speech. By conducting a series of experiments using cross- modal matching tasks, we sought to answer our main question. Baselines were used to establish the validity of our tokens, and then we asked participants to match auditory tokens to the correct visual tokens. Overall, we found that some individuals prioritize emotion information while others prioritize linguistic information when cross-modally integrating speech.

MENTORS

Faculty Mentor: Larry Rosenblum *(right)*
Graduate Student Mentors: James W. Dias, *(lower left)*
Theresa Cook *(lower right)*

Department of Psychology

Abigail Montes De Oca was a Research Assistant at the Riverside Audiovisual Speech and Audition Lab (RASAL) for last year and a half, making her one of the most senior and experienced members of our RA team. From her outstanding scholarship and complete trustworthiness, to her fantastic interpersonal skills, Abby stands out as a leader in every respect. We rely on Abby's intelligence, composure, and diligence to perform the tasks necessary for the daily functioning of our research. From running participants, to producing exacting scientific stimuli, to sensitive data entry, to working with our complex experimental equipment, Abby excels in every area critical to furthering our scientific endeavors. Finally, Abby's great interpersonal skills are especially noteworthy. The most crucial aspect of our experimental enterprise is our engagement with human participants. She understands how essential it is to maintain professionalism, conscientiousness, authority, and compassion. It is something of an art to be both respected and liked by participants, and Abby manages this responsibility with ease. Her manner of interacting with people generates high quality data, and conveys the utmost of ethics and empathy.

AUTHOR

Abigail Montes De Oca

Psychology

Abigail Montes De Oca is a graduating senior in Psychology. She worked as a Research Assistant in Dr. Larry Rosenblum's Riverside Audiovisual Speech and Audition Lab (RASAL) for nearly two years. Her experience with RASAL has positively influenced her in many ways, professionally and personally. She is extremely thankful for the mentorship and support she received from both Theresa Cook and James Dias as it helped her immensely in her academic accomplishments. Abigail also thanks Dr. Rosenblum for the experience provided by working with RASAL, as well as his continued support. She plans to go onto a Master's program in Social Work and, ultimately, into a Clinical Psychology PhD program.

Abigail Montes De Oca

BACKGROUND

In conversation, have you ever felt that a person is listening to the words you are saying but doesn't really grasp the emotion behind your words? Have you had someone completely empathize with the emotions you are portraying but not really hear your words? The ability to identify the emotions of others is vital in all aspects of social interaction. People gain critical information regarding appropriate response through emotion identification and speech.

Past research has shown that both speech and emotion are perceived through more than one sense (multimodally). In other words, people use more than one sensory modality (such as hearing and seeing) to reach an understanding about what a person is saying and feeling. When attending to interpersonal situations, individuals use linguistic and emotional variables to process the information presented. However, it is possible that individuals more readily use either linguistic or emotional information to reach their understanding of the presented situation. For example, when having a conversation, a person may attend more closely to the other individual's expressions or they may pay closer attention to the lexical information they are hearing. In previous research, participants have judged emotional expression using photographs or videos. Studies explored the effects of age, sex and developmental disorder on speech and emotion perception.

Johnson, Emde, Scherer, and Kilnnert (1986) demonstrated the multimodality of emotions. Both visual information and auditory information are essential to identifying emotions; when combined, these cues allow us to accurately interpret the emotions of others. Although it is important to have both aspects to achieve emotion perception, as seen through the studies of Johnson et al. (1986), individuals can still gain some insight into another person's emotion when the attempt is made to separate them. Participants were able to identify some emotions with good accuracy (e.g. sadness) through one modality alone. However, people had trouble identifying other emotions that were easily identified in the visual and auditory combined condition, demonstrating the multimodality of speech perception and emotion.

De Gelder and Vroomen (2000) examined the integration of auditory and visual components in relation to emotion perception and found that audiovisual integration occurs when perceiving emotion. For example, if a participant is shown a happy face paired with a sad voice, he/she is more likely to rate the face as sad. De Gelder & Vroomen (2000) demonstrated that a strong relationship between spoken and visual speech exists when perceiving emotion. Pourtois, de Gelder and Crommelink (2005) studied the relationship between visual and auditory information by analyzing the brain regions activated during emotion perception and found that the middle temporal gyrus was much more activated when participants were shown audio-visual stimuli rather than audio only or visual only stimuli. The higher levels of activation seen in the brain offer evidence that emotion perception is indeed a multimodal process and people attend to both linguistic and emotional cues when detecting emotion.

Researchers have also investigated age effects on multimodal emotion perception and speech. Hunter, Phillips, and MacPherson (2010) tested the emotion detection performance of younger and older individuals when presented with congruent cross-modal stimuli and found that older participants were able to perform just as well as younger participants when identifying congruent faces and voices. However, older participants encountered more difficulty with cross-modally incongruent faces/voices. Although older individuals do not have a more difficult time identifying emotion in a multimodal encounter, they do experience more difficulty when attempting to derive emotional information from only one cue (auditory or visual alone).

Overall, previous research shows the importance of both visual and auditory cues when perceiving emotion. Speech perception is also integrated multimodally; lip-reading has been shown to be a useful tool for better understanding what someone is saying. The multimodality of speech can also be seen through the method of tadoma, a way of feeling speech. In this method, the individual perceives speech by placing his/her hand over the other person's mouth and feeling the ways in which speech is formulated (Rosenblum, 2012). Research has also supported the notion that speech is not only perceived through sound. Dohen, Schwartz, and Bailly

(2010) believe that face-to face interactions offer much more than just sound; the emotions, gestures and facial expressions that are perceived contribute to the perception of speech. We gain vital information about a speaker from interacting with them directly; every part of that interaction is important and gives us a better understanding of the interaction that is occurring.

In the current study, we sought out to see whether individuals gave higher perceptual priority to linguistic or emotional cues in audiovisual speech. In order to establish whether our stimuli were effective, we conducted a series of two baselines, followed by two experiments that used cross-modal matching tasks. In these trials, participants were given two pairs of visual and audio stimuli of a word and the person was to decide which time what they saw best matched to what they heard. For example, a person saw someone say "log" and heard "log," then say saw someone say "log" but heard "dog." The participant had to decide which time the visual and audio stimuli matched.

METHODS AND RESULTS

Method for Baseline 1: Neutral Word ID

In the first experiment, 21 undergraduate students between the ages of 17 and 25 participated; fourteen of the 21 participants were female and seven were male. The students received research credit to fulfill a course requirement. Additionally, the participants had normal or corrected hearing and vision in order to take part in the study. A research assistant had the participants read and sign an informed consent form prior to conducting the experiment. Each participant received detailed directions and was asked about their comprehension of the task to be conducted. The task consisted of matching visual to auditory stimuli. Participants saw/heard two video and audio stimuli pairs; the participant chose the video in which the visual stimuli best matched what they heard. Each participant heard ten words in five pairs. The words differed only on one part (one visible phoneme). For example, the participant saw a model say "camper" while hearing "camper," then saw the same model say "camper" but heard the word "pamper." People had to determine which time what they heard best matched what they saw. It is also important to note that all words were presented neutrally, with no emotion in the face or in the voice. After establishing that the participant fully understood the directions, he/she sat in a sound proof booth to complete the experiment. Directly after the completion of the task, each participant completed a language questionnaire. Overall, completion of the experiment took approximately 15 minutes.

Results for Baseline 1: Word ID

Participants were able to discriminate between incongruent and congruent visual/phoneme tokens at better than chance levels (M correct = 80.9%, SD = 11.0%), $t(19) = 8.97$, $p < .001$). These results indicate that participants could match the word they heard with the word they saw 80.9% of the time, meaning that when participant saw "pamper" and heard "pamper" they were able to accurately pair them, rather than choosing the visual "pamper" with the audio "camper."

Methods for Baseline 2: Emotion Categorization

This experiment sought to test whether participants could discriminate three distinct emotions in the stimuli: happy, mad, or sad. Participants experienced three presentations: participants saw and heard audio only, visual only, or audio-visual stimuli. Ten (6 female, 4 male) UCR undergraduate psychology students between the ages of 17 and 21, who had not participated in the neutral baseline, participated in exchange for research credit. As in the Neutral Baseline, research assistants followed the same general protocol. However, participants now categorized stimuli based on emotion. They heard the same ten words as in the first experiment, but in three distinct emotions: happy, mad, or sad. Participants were to categorize each auditory, visual or audiovisual stimulus into emotional categories. The entire procedure for running this experiment remained very similar to Experiment 1, and it took approximately 15 minutes for each participant to complete.

Results for Baseline 2: Emotion Categorization

Experiment 2 contained three conditions: audio only (AO), video only (VO) or audio-visual (AV). Participants were able to successfully categorize emotion in all conditions:

AO (M = 84.2%, SD = 9.0%), VO (M = 90.8%, SD = 3.2%), and AV (M = 94.0%, SD = 3.5%). In addition, the three conditions significantly differed from one another: AO versus VO, t (9) = 2.8, p = .021; VO versus AV, t (9) = 2.7, p = .025; and AO versus AV, t (9) = 4.2, p = .002. In other words, participants were able to categorize the emotion they saw and heard best, with an accuracy level of 94%. Participants also fared well at categorizing emotions in the face they saw (correct 90.8% of the time). Additionally, they were also able to categorize emotion from the voice they heard with excellent accuracy (they correctly identified the emotion in the voice they heard 84.2% of the time). All three conditions significantly differed from chance, with the least difference in the AO condition, t (9) = 29.8, p < .001.

After determining that our tokens were both linguistically and emotionally discriminable, we conducted our experiments.

Methods for Experiment 1: Combined Emotional and Linguistic Cross-modal Matching

Participants (N= 20, 12 female, 8 male) for this study were also undergraduates at the University of California, Riverside and were between the ages of 17 and 23. We tested whether participants were better able to distinguish fully linguistically and emotionally congruent audiovisual in information speech from speech which was incongruent on only one of those factors. In order to test this question, participants saw and heard fully congruent audio-visual stimuli and partially incongruent (linguistically or emotionally, but not both) audio-visual stimuli. The experiment contained two types of trials: In one type of trial, participants compared fully congruent (FC) stimuli to linguistically incongruent but emotionally congruent (EC) stimuli. In the second type of trial participants compared fully congruent (FC) stimuli to those which were emotionally incongruent but linguistically congruent (LC). For example, in FC versus EC trials, participants heard a happy voice say "pamper" and saw a happy person say "pamper" in one stimulus, then in the second stimulus they heard a happy voice say "pamper" but saw a happy person say "camper." In the FC versus LC trials, participants heard a happy voice say "pamper" paired with a happy face saying "pamper" in one stimulus, while in the second stimulus they heard a happy voice say "pamper" but saw a sad person say "pamper." In each of the trials, the participant was to best match what they saw with what they heard. Two distinct actors were used, one male and one female. Participants either judged stimuli containing the female actor or the male actor to limit the number of trials. As in the other experiments, the only procedural change was the explicit directions. Figure 1 further clarifies the method used in each phase.

Results for Experiment 1: Combined Emotional and Linguistic Cross-modal Matching

We found no difference in participants' responses based on seeing and hearing the male or female actor, F(1,18) = 1.289, p = .264. We found no difference in people's ability to distinguish FC stimuli from EC or LC stimuli, F(1,18) = .560, p = .459. People were just as good at discriminating fully congruent from partially congruent stimuli when the incongruency was emotional as when the incongruency was linguistic. Interestingly, an interaction was found in regards to actor and information type (linguistic/emotional), F(1,18) = 6.118, p = .018. Emotionally congruent information for the male model was more easily matched (M = 82.7%, SD = 16.8%) than for the female model (M = 67.5%, SD = 11.4%), t(18) = 2.69, p = .015. However, this was not the case concerning linguistic information: The female model's (M = 74.8%, SD = 9.5%) linguistic information was not more easily matched than the males model's (M = 69.2%, SD = 14.3%), t(18) = 1.478, p = .157. Participants performed the discriminating task of fully congruent from partially congruent (linguistic or emotional) at higher than chance levels, t(19) = 9.292, p < .001.

Methods for Experiment 2: Emotional versus Linguistic Salience Preference

After establishing the validity and discriminability of our tokens, we conducted a second experiment to attend to the main question: whether people perceptually prioritize emotional versus linguistic information when perceiving audio-visual speech. Participants were twenty undergraduate students enrolled in introductory psychology courses at the University of California, Riverside.

1. [**FC**] Fully audiovisually congruent stimuli.		
Example: Participants heard "camper" in a happy voice and saw "camper" articulated with a happy facial expression.	**"CAMPER"** ☺	 (CAMPER)
2. [**EC**] Emotionally congruent and linguistically incongruent stimuli.		
Example: Participants heard "camper" in a happy voice and saw "pamper" articulated with a happy facial expression.	**"CAMPER"** ☺	 (PAMPER)
3. [**LC**] Linguistically congruent and emotionally incongruent stimuli.		
Example: Participants heard "camper" in a happy voice and saw "camper" articulated with a sad facial expression.	**"CAMPER"** ☺	 (CAMPER)
4. [**FI**] Emotionally and linguistically incongruent audiovisual stimuli.		
Example: Participants heard "camper" in a happy voice and saw "pamper" articulated with a sad facial expression.	**"CAMPER"** ☺	 (PAMPER)

Figure 1. *In Experiment 1, participants saw and heard fully congruent audio-visual stimuli and partially incongruent (linguistically or emotionally, but not both) audio-visual stimuli. The experiment contained two types of trials: In one type of trial, participants compared fully congruent (FC) stimuli to linguistically incongruent but emotionally congruent (EC) stimuli, and in the other trial participants compared fully congruent (FC) stimuli to those which were emotionally incongruent but linguistically congruent (LC). In Experiment 2, on half of the trials participants compared one emotionally congruent/linguistically incongruent (EC) stimulus to one linguistically congruent/emotionally incongruent stimulus and judged which was the best audiovisual match. On the other half of trials, participants saw one fully congruent (FC) and one fully incongruent (FI) stimulus and judged which was the best audiovisual match. These trials served as catch trials to ensure that participants were attending to the task. Emoticons represent tone of voice in auditory portion of stimuli.*

Experiment 2 was to test whether participants favored emotional content or linguistic content. In this experiment, on half of the trials participants compared one emotionally congruent/linguistically incongruent (EC) stimulus to one linguistically congruent/emotionally incongruent stimulus and judged which the best audiovisual match was. For example, the person might hear a happy voice saying "camper" while seeing a happy face say, "pamper," (EC) then the participant might hear a happy face say "pamper" but hear a sad voice say "pamper" (LC). The participant judged which time what they heard best matched with what they saw. On the other half of trials, participants saw one fully congruent (FC) and one fully incongruent (FI) stimulus and judged which the best audiovisual match was. These trials served as catch trials to ensure that participants were attending to the task. The procedure for this experiment was very similar to that of the previous studies.

Results for Experiment 2: Emotional versus Linguistic Salience Preference

Once again in the FC/FI trials, participants selected completely congruent stimuli at greater than chance levels $t(19)= 45.366$ $p<.001$. Distinct preferences for selecting emotional (M= 73.6% SD= 11.1) or linguistic (M=75.1% SD= 10.3) content as the best audiovisual match were displayed by the participants in the EC/LC trials. Eleven participants judged EC stimuli as the best audiovisual match 73.6% of the time, while nine participants judged LC stimuli as the best audiovisual match 75.1% of the time.

CONCLUSION

Although all stimuli proved to be highly discriminable, further research was needed to truly understand if individuals perceptually prioritize linguistic or emotional information in a cross-modal matching task. Experiment 2 included tokens using the both the male and female models. Additionally, Experiment 2 tested whether participants perceptually prioritize information in cross-modal matching. Through the analysis of Experiment 2, we found that some people highly prioritize emotional versus linguistic content when matching auditory to visual speech. Of the twenty participants in Experiment 2, nine of them selected linguistic content, while eleven selected emotional content as the best match. It would be interesting for future studies to explore why some individuals prefer emotional versus linguistic content and what determines this preference. Future studies should also compare scores from the beginning of the task to those at the end, in order to see whether participants change their preferences.

Our findings allow us to further understand the multimodality of speech and emotion by showing that individuals attend to information differently. Some individuals prioritize linguistic information and others prioritize emotional information; utilizing this concept we can develop new teaching tools that are directed to each type of person. For example, individuals can be taught to attend to either emotion or linguistic content; this can be useful for those with impairment.

DISCUSSION

Now that the question of whether individuals perceptually prioritize information in cross-modal speech matching has been addressed, it would be beneficial to conduct further experiments using visually or hearing impaired participants. Some research has already been conducted using such populations. For example, Dyck, Farrugia, Shochet, and Holmes-Brown (2004) analyzed emotion recognition in children with sight or hearing impairments. Results indicated that visually impaired and hearing-impaired groups showed signs of deficit regarding emotion recognition (Dyck et al., 2004). Ludlow, Heaton, Rosset, Hills, and Deruelle (2010) also investigated the emotion perception of deaf children and found that normal hearing children could identify emotion more easily than those with hearing impairment of speech and emotion perception despite modality. It would be beneficial to know whether those with impairment perceptually prioritize linguistic or emotional information and how that compares to individuals without impairment.

REFERENCES

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

Dohen, M., Schwartz J., & Bailly G. (2010). Speech and face-to-face communication — An introduction. *Speech Communication,* 52(6), 477-480.

Dyck, M J., Farrugia C., Shochet I., & Holmes-Brown M. (2004). Emotion Recognition/ understanding ability in hearing or vision-impaired children: Do sounds, sights, or words make the difference? *Journal of child Psychology and Psychiatry and Allied Disciplines*, 45, 789-800.

Hunter, E M., Phillips L., & MacPherson S. (2010). Effects of age on cross- modal emotion perception. *Psychology and Aging,* 25, 779-787.

Johnson W., Emde N., Scherer K., & Kilnnert M. (1986). *Recognition of Emotion from Vocal Cues*, 43, 280-283.

Ludlow, A., Heaton P., Rosset D., Hills P., & Deruelle C. (2010). Emotion recognition in children with profound and severe deafness: Do they have a deficit in perceptual processing? Neuropsychology, development, and cognition. *Section A, Journal of Clinical and Experimental Neuropsychology*, 32, 923-928.

Pourtois, G., de Gelder, B., & Crommelink M. (2005). Perception of facial expressions and voices and of their combination in the human brain. *Cortex,* 41, 49-59.

Rosenblum, L., Speech [PDF]. Retrived from online lecture notes site : http://ilearn.ucr.edu/webapps/portal/frameset.jsp?tab_tab_group_id=_2_1&url=%2Fwebapps%2Fblackboard%2Fexecute%2Flauncher%3Ftype%3DCourse%26id%3D_108685_1%26.