Western Undergraduate Psychology Journal

Volume 5 | Issue 1 Article 3

2017

The Recognition of Vocal Expression of Emotion in Children

Andrea Jennings
ajenni2@uwo.ca

The Recognition of Vocal Expression of Emotion in Children

Andrea Jennings*

This study examined the effect of emotional and verbal congruency on children's judgements of emotion. Children listened to happy and sad verbal phrases with congruent or incongruent prosody, or congruent or incongruent background music, then rated the emotion expressed in the recordings on a five-point happy face scale. Results revealed that for happy literal phrases, congruent phrases were rated significantly higher than incongruent phrases, $F(1, 12) = 35.15$, $p < 0.001$, partial $n2 = 0.75$. For sad literal phrases, congruent phrases were rated significantly lower than incongruent phrases, $F(1, 12) = 6.23$, $p < 0.03$, partial $n2 = 0.34$. The effect of congruency on ratings of both phrase types demonstrates that children as young as 3-years-old are able to use external cues and to infer underlying emotions based on literal meanings to make judgements about emotion. The differential uses of the various cues to emotion by children is likely due to their experience with different emotions, and the salience of cues within happy and sad intonation and music.

Many studies have examined children's ability to visually recognize emotions through facial expressions, but few have observed their ability to recognize verbal portrayals of emotion. Children's recognition of verbal emotion is a valuable field of study as results can be used by adults to express verbal emotion in ways that are the easiest for children to decipher. The current study aims to determine the modality, in other words, the nature of the emotional information, that best supports vocal emotion recognition. Emotional information can be expressed in several different ways, such as the actual literal or semantic meaning of the words used (e.g., "I am sad"). Alternatively, vocal prosody may be used to convey emotion through the tone of one's voice. Finally, the portrayal of emotion may be affected by the presence of background music (e.g. sad music playing during a sad scene in a movie).

It has been found that the addition of matching, or congruent, cues to emotion helps adults with emotion detection. Thompson, Russo, and Quinto (2008) conducted a study to determine if visual information was integrated with audio information when detecting the emotion of a sung phrase in adults. The pairings of audio and video were either incongruent (happy audio with a sad video or sad audio with a happy video) or congruent (happy audio with a happy video or sad audio with a sad video). Participants made emotional ratings of sung intervals in the single-task condition, and rated the emotion of the sung intervals while performing a secondary task in the dual-task condition. For both conditions, emotional ratings were higher when the emotion in the video and the audio matched, suggesting that participants used both visual and acoustic cues when determining emotion (Thompson et al., 2008). While this study exemplifies the mixed use of visual and acoustic cues, few studies have explored the varying roles of different modalities of acoustic cues.

In addition to the ability to recognize emotion through music, adult listeners can detect emotion in both familiar and unfamiliar instrumental music (Balkwill, Thompson, & Matsunaga, 2004). In Balkwill, Thompson, and Matsunaga's (2004) study, Japanese participants rated Japanese, Western, and Hindustani music

Author Note: Thank you to Dr. Christine Tsang, Chair of the Department of Psychology at Huron University College, the parents and families of the participants, and the YMCA of Western Ontario for their support of this research. Manuscript is based on data used in an independent study course.

*Initially submitted for Psychology 3998F and 3999G at Huron University College at the University of Western Ontario. For inquiries regarding the article, please email the author at *ajenni2@uwo.ca*.

for the emotions joy, anger, and sadness. No significant difference in ratings was found between culturally familiar and culturally unfamiliar instrumental music, suggesting that the emotional basis of music is portrayed through acoustic cues in which familiarity is not necessary (Balkwill et al., 2004). This suggests that the emotional basis of music is not based on experience with the particular song, but on the universal understanding of acoustic features.

Although the ways in which acoustic features portray emotion may be universal, the recognition and interpretation of this emotion can be impacted by training experience. In a study of the effects of training experience, Mualem and Lavidor (2015) focused on a short-term music intervention that aimed to determine the ways that music conveys emotion. Young adults attended four 30-minute music intervention sessions in which they discussed emotions and how they are portrayed in music. Results demonstrated that the intervention significantly increased the accuracy of emotional recognition in young adults (Mualem & Lavidor, 2015). These findings can be extended to determine if the presence of background music also helps the vocal emotion recognition process. Mualem and Lavidor's (2015) results also suggest that phrases with emotional prosody and phrases with background music will have similar emotion ratings, as acoustic and linguistic cues are often similar. The notion that music intervention increases the accuracy of emotional recognition demonstrates that background music may be an effective means of portraying emotion. As young adults were shown to utilize cues from music to assist in the emotion-detection process, there may also be an impact of background music on children's judgements of vocal emotion.

In order for participants to be able to use background music and intonation to assist in the detection of emotion, they must be able to attend to two acoustic signals at once. In a study by Russo and Pichora-Fuller (2008), it was found that word identification was affected by background music for younger adults, but not for older adults. Russo and Pichora-Fuller (2008) found that when attempting to identify words, younger listeners (aged 18-30 years) paid attention to both the speech and music, while older listeners (aged 65-78 years) focused on the speech only. The results of this study indicate that background music and words may be used independently by young adults to determine emotion.

To our current knowledge, no studies have directly examined children's ability to verbally recognize emotions through different modalities. The current study will determine which cue for emotion children rely on the most: literal meaning, vocal prosody, or background music. A significant effect of congruency or stimulus type may suggest that children use more resources than adults when attempting to recognize speech, due to their lack of experience in recognizing the emotions of others. As highlighted by Russo and Pichora-Fuller (2008), younger adults pay attention to all acoustic cues, while older adults only rely on speech itself. Utilizing children in a study of emotion detection will demonstrate what cues people use while they are in the beginning stages of learning to detect and interpret emotions. The results could have implications in childhood education settings, and provide evidence for the best way to communicate emotions to children. The current study has an exploratory purpose to determine the modality (literal meaning, vocal prosody, or background music) children rely on the most when interpreting vocal emotion. Due to the work by Russo and Pichora-Fuller (2008), emotional ratings are expected to be the most dependent on background music.

Method

Pre-Test Phase

Fifteen short phrases derived from 12 different books written for children were used to retrieve five happy, five neutral, and five sad literal phrases. A female voice read each phrase monotonously. A Yeti USB microphone was used to record the stimuli.

In the pre-test phase, phrase stimuli included recordings of monotonous happy literal phrases, monotonous neutral literal phrases, and monotonous sad literal phrases. Prosody stimuli included each of the 15 literal phrases expressed with both happy and sad intonations. Background music stimuli included happy instrumental music, neutral instrumental music, and sad instrumental music.

Throughout the pre-test phase, a group of 10 adults heard recordings of monotonous phrases, phrases with prosody, and excerpts of background music. First, monotonous recordings of the five phrases with happy literal content, five phrases with neutral literal content, and five phrases with sad literal content were played. Participants rated the phrases on a happy face scale of one (sad) - five (happy) by circling one of the faces. Next, participants rated five phrases with happy prosody, five phrases with neutral prosody, and five phrases with sad prosody on the happy face scale. Finally, participants rated five excerpts of happy music, five excerpts of neutral music, and five excerpts of sad music on the happy face scale. All recordings were played in a quasi-random order to ensure no two stimuli of the same emotion were played back to back. The most constant nine monotone phrases (three happy, three neutral, and three sad), nine vocal prosodies (three happy, three neutral, and three sad), and nine music excerpts (three happy, three neutral, and three sad) were selected to be used as the stimuli for the child test phase.

Child-Test Phase

Participants.

Thirteen 3 to 9-year-old children were recruited from a YMCA Child Care Center in London, Ontario, and from a developmental participant list maintained by the Western Department of Psychology. Five females and eight males participated.

Materials.

The stimuli used by the child participants were selected from the stimuli with the highest emotional ratings from the pre-test phase. For example, a “happy” stimulus would have a mean rating by adult listeners of near five whereas a “sad stimulus” would have a rating near one. Stimuli were congruent, incongruent, or neutral. Congruent stimuli included phrases in which the literal meaning matched the vocal prosody or music. The congruent stimuli were pairings of happy literal phrase and happy voice, happy literal phrase and happy music, sad literal phrase and sad voice, and sad literal phrase and sad music. Incongruent stimuli included phrases in which the literal meaning was mismatched with the vocal prosody or music. The incongruent stimuli were pairings of sad literal phrase and happy voice, sad literal phrase and happy music, happy literal phrase and sad voice, and happy literal phrase and sad music. Neutral phrases were phrases in which the literal meaning was neutral as rated by adults in the pre-test phase. Neutral stimuli were neutral phrases matched with neutral prosody or neutral music. The music recordings were controlled for percussion, as sad music typically contains less percussion than happy music, which could have had an impact on ratings. Children rated the recordings by colouring the faces on a five-point happy face scale, which was consisted of five faces (frowning, neutral, and smiling) along a continuous line.

Procedure.

In the child test phase, children were introduced to the experimenters and were presented with the visually-based happy face scale. Before the testing began, participants were asked to point to the sad face and the happy face to ensure that the child understood how to use the scale accurately. In a quasi-random order, children listened to recordings of the congruent, incongruent, and neutral stimuli. A total of six monotonous literal phrases (two happy, two neutral, two sad), six excerpts of music (two happy, two neutral, two sad), four congruent phrases with prosody (two happy, two sad), four incongruent phrases with prosody (two happy, two sad), six congruent phrases with music (two happy, two neutral, two sad),

and four incongruent phrases with music (two happy, two sad) were played for the children aged four-years-old and greater. Three-year-old children listened to three monotonous literal phrases (one happy, one neutral, one sad), three excerpts of music (one happy, one neutral, one sad), two congruent phrases with prosody (one happy, one sad), two incongruent phrases with prosody (one happy, one sad), five congruent phrases with music (two happy, one neutral, two sad), and four incongruent phrases with music (two happy, two sad).

After each recording, the participant was instructed to colour one face on the happy face scale. They were instructed to colour the happy face, neutral face, sad face, or somewhere in between, depending on what emotion they thought was expressed in the recording. Each trial took approximately 30 seconds to play the recording and obtain a rating. Participants who were 4-years-old and greater completed 30 trials each, for a total participation time of approximately 25 minutes. Three-year-old participants completed 21 trials each for a total participation time of approximately 20 minutes. Three-year-old participants completed fewer trials due to their significantly shorter attention spans than the older children.

Each face on the visual scale was assigned a number between one (sad) and five (happy). Total scores were calculated for each stimulus type.

Results

Preliminary Analyses

Paired samples t-tests were conducted to determine if the children were able to detect differences between the monotonous happy and sad literal phrases, happy and sad prosody, and happy and sad music. Results revealed that monotonous happy literal phrases ($M = 4.23$, $SD = 1.20$) were rated significantly higher than monotonous sad literal phrases ($M = 2.00$, $SD = 1.24$), $t(12) = 3.70$, $p < 0.003$, $d = 1.03$, happy prosody ($M = 4.65$, $SD = 0.85$) was rated significantly higher than sad prosody ($M = 1.85$, $SD = 1.07$), $t(12) = 5.53$, $p < 0.004$, $d = 1.54$, and happy music ($M = 4.15$, $SD = 1.03$) was rated significantly higher than sad music ($M = 3.00$, $SD = 0.87$), $t(12) = 3.97$, $p < 0.001$, $d = 1.10$. Parametric analyses were conducted separately for happy literal phrases (e.g. Happy birthday, David M! We had a party at school with cupcakes) and for sad literal phrases (e.g. Mommy, Mommy! Do something about this tooth. It hurts so much I can't even eat my apple!). Analyses were conducted separately for happy literal phrases and sad literal phrases to determine if there was a difference in the effect of congruency and stimulus type on happy versus sad literal phrases.

Happy-Literal Phrases Analyses

A two-way congruency x stimulus type within-subjects analysis of variance revealed that ratings were significantly different for congruency and stimulus types of happy literal phrases. Congruent phrases (happy literal phrase and happy prosody or happy music) ($M = 4.46$, $SD = 1.10$) were rated significantly higher than incongruent phrases (happy literal phrase and sad prosody or sad music) ($M = 2.90$, $SD = 1.50$), $F(1, 12) = 35.15$, $p < 0.001$, partial $n2 = 0.75$. There was also a significant difference between stimulus types, with music having a greater impact on ratings than prosody, $F(1, 12) = 9.11$, $p < 0.02$, partial $n2 = 0.43$.

Sad-Literal Phrases Analyses

A two-way congruency x stimulus type within-subjects analysis of variance revealed that ratings were significantly different for congruency, but not for stimulus type of sad literal phrases. Congruent phrases (sad literal phrase and sad prosody or sad music) ($M = 1.83$, $SD = 0.90$) were rated as significantly lower than incongruent phrases (sad literal phrase and happy prosody or happy music) ($M = 2.24$, $SD = 1.20$), $F(1, 12) = 6.23$, $p < 0.03$, partial $n2 = 0.34$. There was no significant difference in ratings between stimulus types, $F(1, 12) = 0.88$, $p > 0.05$.

Discussion

There is extensive research demonstrating children's visual emotion recognition, but there is a lack of literature regarding verbal emotion recognition. Overall, results of the current study indicated that congruency had an impact on verbal emotion recognition for both happy and sad literal phrases, but an effect of stimulus type was only found for happy literal phrases. Congruent phrases were rated as more extreme (higher for happy literal phrases and lower for sad literal phrases), and music had a larger impact on ratings than prosody for happy literal phrases, as shown by the significant main effect of stimulus type on ratings for happy literal phrases. There was no significant difference of impact on ratings between intonation and music for sad literal phrases.

Previous findings by Russo and Pichora-Fuller (2008), showed that younger adults paid attention to both speech and music when attempting to identify words, but older listeners focused on the speech only. The present study extends these findings by demonstrating that 3 to 9-year-old children attend to semantic, prosodic, and musical cues when attempting to identify the underlying emotion of a verbal utterance. As ratings were significantly different between congruent and incongruent phrases for both happy and sad literal phrases, children not only attended to, but also utilized all cues towards emotion when making their emotional ratings.

The effect of congruency on ratings on both phrase types demonstrates that children as young as 3-years-old are able to use external cues and to infer underlying emotions based on literal meanings to make judgements about emotion. Congruent cues allowed the children to combine their understandings of the underlying emotions of the various cues. When the cues conflicted in emotion, the incongruent background music and intonation made the emotion detection task more difficult and greatly increased cognitive load, which resulted in a significant difference in ratings.

Interestingly, happy literal phrases were rated as sad when they were paired with sad music ($M = 2.23$), but sad literal phrases were not rated as happy when paired with happy music ($M = 2.12$). The fact that the sad music had a larger effect than the happy music when paired with an incongruent phrase indicates that the emotional cues within the sad music were more obvious than those within the happy music, or that the sad literal phrases were more clearly sad than the happy phrases were happy. The sad literal phrases included words that could have been more clearly sad than the happy words were happy, such as "cry" and "hurt". In contrast, participants may have had to use more inference when detecting the emotions of the happy literal phrases, as the happy emotion was less obvious. Consequently, the happy literal phrases, which were not as clearly happy as the sad phrases were sad, were more affected by the external cues of intonation and music.

The significant difference between stimulus type for happy literal phrases is presumably due to the salience of the acoustic cues within happy music in comparison to the cues of happy intonation. As highlighted by Mualem and Lavidor (2015), acoustic cues and linguistic cues are often acoustically similar. Music typically has more exaggerated intonation and acoustic patterns than prosodic speech, resulting in easier emotion detection. Due to this, the children in the current study may have been able to use background music to help detect happy emotion more efficiently than they were able to use intonation. Furthermore, the majority of children's music has a fast tempo and is high pitched, similar to the happy music used in the current study. This prominence of happy music in a child's life provides them with ample experience to detect and recognize happy music.

The nonsignificant difference between stimulus types for sad literal phrases is hypothesized to be due to children's lack of experience with sad emotions. The typical child is exposed to happy speech, phrases, and music relatively consistently throughout their development (Ziv & Goshen, 2006). Over time,

a child practices deciphering these linguistic cues, and is usually positively reinforced when they make an appropriate response. Due to this, children are fairly good at recognizing happy phrases, intonation, and music. Although children may be exposed to some sad speech, phrases, and music, these occurrences are few and far between in comparison to those that are happy (Ziv & Goshen, 2006). Due to a lack of practice in determining the underlying emotion behind sad literal phrases, participants utilized all the cues that were available to interpret the underlying emotion.

The location that participants completed the study may have had an effect on their ratings. While the 3-year-olds were in a daycare setting, the older children completed the study in a lab room. The daycare setting included background noise from other children and visual distractions from toys and activities in the room. In the lab, participants had few distractions, and worked one-on-one with the experimenter. Therefore, participants in the lab were presumably able to focus on the task more efficiently, which may have resulted in more accurate ratings. Additionally, this study could have had a stronger reliability if a larger population was observed. Flawed results may have resulted due to the small number of children observed. A larger sample population would also allow researchers to examine the effects of age on recording ratings. Furthermore, the use of a female voice for the recordings may have had an effect on emotional ratings. Emotional cues in a female voice may be more or less apparent than those in a male voice. Future studies should utilize a larger sample size, and may also analyze the effects of participant age and the gender of the voice recording on emotional ratings.

Adults were utilized in the pre-test phase to determine the most constantly rated and therefore most salient emotional stimuli. Although this was a study of the verbal emotion recognition in children, adults were used to obtain the stimuli for the child-test phase as adults are more experienced at emotion detection and recognition, and therefore were presumably more accurate at detecting the emotions within the stimuli than children would have been.

Furthermore, future studies should determine if the child's emotional state at the beginning of the test phase has an impact on their ability to recognize vocal emotion. If a child is in a sad mood at the beginning of the study, their current emotional state may prime them to better recognize sad emotions in all modalities. Consequent studies should measure the child's emotional state prior to completing the ratings, and determine if there is a significant effect of prior emotional state on ratings. Additionally, the current study could be replicated with a longitudinal design, in which each child would complete the ratings multiple times across a set time frame. If the same child completes the study over separate days, weeks, or even months, their ratings may differ over time. A replication of the current study using a longitudinal design could also measure participants' personality development to determine if a change in ratings of emotions would be due to a change in the child's overall personality or due to a change in their emotion recognition abilities.

The current study did not examine the impact of age on ratings or cues utilized to detect emotions, and the ages of the participants ranged greatly. As a result, the older children's ratings and the younger children's ratings may have unequally affected the overall results. At a young age, children need to use various cues in order to make accurate judgements about the content of an utterance (Hoff, 2014). As children's language abilities develop and they gain experience with different emotions, this necessity may diminish, as they become better able to break down the words of utterances and make assumptions about what the words mean when put together. Future studies should include a smaller age range in order to minimize any confounding variables. Furthermore, future studies should examine the variability of the congruency effect among children of different ages. It is hypothesized that the effects of ambiguity on verbal emotion recognition will

differ throughout the lifespan. At different ages, children likely utilize different cues to emotion. As children gain more experience understanding the literal meaning of phrases, the external cues of intonation and background music are expected to have a smaller impact on emotion recognition. Additionally, examining congruency effects of emotional phrases with pre-linguistic participants could prove to be valuable. Future studies could use head-turn preference procedures to determine if pre-linguistic infants and toddlers are able to recognize when emotional cues conflict. Results could demonstrate when children start to recognize and understand various acoustic cues.

Ultimately, the results of the current study are valuable for both home and educational settings to assist children in verbal emotion recognition. As the results show that intonation and background music both have an impact on emotion detection, it is evident that children as young as three years old are able to integrate various cues into their understanding of emotions. Results of this study could be especially useful for educators of children to assist with emotion recognition. Providing various congruent and clear cues could prove extremely useful in helping children distinguish emotions. Furthermore, conflicting and incongruent background sounds and music should be avoided as children clearly pay attention to all of the cues that are available.

The ability to recognize verbally portrayed emotion is extremely important for detecting the intentions of others (Krothapalli & Koolagudi, 2013). As emotion recognition is a significant predictor of social development (Williams & Gray, 2013), greater focus needs to be placed on assisting children in the recognition of verbal emotions. The results of the current study demonstrate the cues children utilize when they are in the beginning stages of learning to detect and interpret emotions. This knowledge will allow for adults and childhood educators to communicate emotions in ways that are the most easily interpretable by children.

First Received: 12/19/2016
Final Revision Received: 05/12/2017

References

Balkwill, L., Thompson, W., & Matsunaga, R. (2004). Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners. Japanese Psychological Research, 46(4), 337-349.

Hoff, E. (2014). Language development (5th ed.). Belmont, CA: Wadsworth Cengage Learning

Krothapalli, S. R. & Koolagudi, S. G. (2013). Emotion Recognition Using Speech Features. New York: Springer.

Mualem, O., & Lavidor, M. (2015). Music education intervention improves vocal emotion recognition. International Journal of Music Education, 1-13.

Russo, F., & Pichora-Fuller, M. (2008). Tune in or tune out: Age-related differences in listening to speech in music. Ear and Hearing, 29(5), 746-760.

Thompson, W., Russo, F., & Quinto, L. (2008). Audio-visual integration of emotional cues in song. Cognition & Emotion, 22(8), 1457-1470.

Williams, B., & Gray, K. (2013). The relationship between emotion recognition ability and social skills in young children with autism. Autism, 17(6), 762-768.

Ziv, N., & Goshen, M. (2006). The effect of ‘sad’ and ‘happy’ background music on the interpretation of a story in 5 to 6-year-old children. The British Journal of Music Education, 23(3), 303-314.