Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Effect of Age and Gender on Categorical Vocal Emotion Recognition in Mandarin Chinese

Yu Chen[1], *Ting Wang*[2,3], *Hongwei Ding*[1*]

[1]Speech-Language-Hearing Center, School of Foreign Languages,
Shanghai Jiao Tong University, China
[2]School of Foreign Languages, Tongji University, China
[3]Center for Speech and Language Processing, Tongji University, China

garce_chan@sjtu.edu.cn, 2011ting_wang@tongji.edu.cn, hwding@sjtu.edu.cn

Abstract

Categorical perception (CP) effect, as a fundamental feature of perception which allows humans to rapidly and properly respond to sensory cues, has received accumulating evidence from various perceptual experiments, such as musical tone perception and facial emotion identification. The current study attempted to detect the potential CP effect of vocal emotion perception in Mandarin Chinese and further explored the effect of age and gender on emotional prosody categorization. Three emotional prosody continua (happiness-sadness, neutral-happiness, neutral-sadness) were created using speech synthesis. Each continuum consists of 11 emotional speeches with equal-sized physical differences. Two groups of subjects (children and adults) were instructed to finish both the identification tasks and discrimination tasks. Results indicated that adults showed significantly narrower boundary width compared to younger children, and women's boundary width was also significantly narrower than men's. Moreover, the emotion categories of the children group did not coincide with those of adults. These findings provide novel evidence for categorical emotion recognition from voice and also illuminate the effect of age and gender on prosodic perception of emotional speech. Particularly, the current research indicates that children's capacity to decode emotions follows a slow course of development before changing into adults' emotional category patterns.

Index Terms: emotional prosody, categorical perception, Mandarin Chinese, age, gender

1. Introduction

Categorical perception (CP) has been widely investigated since the concept was put forward and explained systematically by Liberman et al. [1]. CP effect occurs when continuous sensory inputs are perceived and coded as discrete categories by the human brain [2, 3, 4]. Equal-sized physical differences between stimuli can be perceived as larger or smaller depending on whether the stimuli are within the same category or not [3]. This does not mean that individuals cannot distinguish the nuance between different stimuli, but signifies that some sensory differences are more meaningful and significant than others.

Whether vocal expressions are perceived as varying continuously along underlying dimensions [5] or as belonging to qualitatively discrete emotion categories [6] is still a matter of debate. Investigating CP of vocal expressions could be of great theoretical importance, because evidence of CP would reconcile with a discrete-emotions framework instead of a dimensional model [7]. In the past two decades, many studies have shown that the perception of facial emotion fits the criteria of CP. However, the expressing and perceiving of emotional signals is not limited to the visual channel, it was also related to vocal channel. Vocal perception is important in comprehending the speaker's emotions, since it is more independent of speaker distance. Every emotion has its unique characteristics in the acoustic aspect that differentiate it from other emotions [8, 9, 10, 11]. And the unique patterns for each emotion (at least those which were classified as basic emotions) provide a premise for the CP effect in emotional perception.

Previous studies have done massive work on age effect in emotional perception [12, 13]. It has come to a consensus that the child's capacity to comprehend and accurately decode facial emotion follows a slow developmental course [14]. Additionally, it has previously been observed that gender differences exist in the emotional sensory process. Women tend to show better performance than men in identifying the meanings of nonverbal cues of face, body, and voice [15,16], and this superior ability strengthens women's advantage in perceiving speech emotions. Massive studies have detected age effect and gender difference from visual channel [17, 18]. In light of previous findings, we predict that age effect and gender difference might also occur in the perception of vocal emotions. The current study attempts to investigate identification and discrimination performance in children and adults to test whether they show categorical perception of vocal emotional expressions (happiness and sadness) and further investigate the potential age effect and gender difference on Mandarin speakers' auditory perception of emotional prosody.

2. Methods

2.1. Participants

Two groups of people were recruited as the participants. 24 children were recruited from primary schools, including 14 boys and 10 girls. The mean age was 8.58 (*SD* = 1.67). 32 university students were recruited as the adult group, including 16 males and 16 females. The mean age was 24.84 (*SD* = 1.17). All the participants are native Mandarin speakers and had neither listening loss nor intellectual or cognitive difficulties.

* Corresponding author

10.21437/SpeechProsody.2022-118

2.2. Stimuli

Two basic emotions, happiness and sadness, were selected in this experiment. They are the two most common emotions in daily life and differ in physiological arousal [8]. Neutral emotional stimuli were also included as the filler. In the following experiment, true disyllabic Mandarin word “森林”, which means ‘forest’ in English, was chosen because of its higher identification rates in all the three emotions in a screening test before. The word was recorded by native speakers who had acting experience. The sampling rate and sampling frequency were 44.1KHz and 16 bits, respectively.

The stimuli used in this perception task were synthesized by a script in Praat. The script copied prosody (syllable duration and pitch) from source to target with the same step size, as a mixture of source’s and target’s durations and pitch points. The values of synthesized acoustic cues were linearly interpolated between the values of the source stimulus and the target stimulus. The first and the last steps were equivalent to the source and target prototype respectively. The number of steps was 10. Morphs were created in proportions of 10:90 (e.g., for the neutral to happiness continuum, 10% neutral and 90% happiness, which marked as N1H9 in the current study), 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, 90:10. Therefore, three continua, N-H (neutral-happiness), N-S (neutral-sadness), H-S (happiness-sadness) were included, with each continuum containing 11 vocal stimuli. The mean *F0* of happiness, sadness, neutral prototype is 382.6, 224.9, 241.5 Hz, respectively.

2.3. Experimental design

In the present study, identification and discrimination tasks were adopted to investigate potential CP performance in perception of the three continua [4, 19, 20, 21]. In identification task, the participants were required to choose the emotion of each synthesized stimuli by forced choice. The stimuli within each continuum were played twice randomly. In discrimination task, the participants were required to differentiate between pairs of synthesized stimuli and judge whether the pairs of stimuli were the same or not. The inter-stimulus interval within each pair was 500 ms. The stimuli in each pair consisted of two stimuli which belonged to the same emotional continuum but differed by 20%. Each stimuli pair was played twice in different presentation orders. Five pairs of identical stimuli were also included as filters in the discrimination task.

2.4. Data analysis

To investigate the potential categorization phenomenon in emotional prosody, Probit analysis [22] was conducted to assess the boundary position and boundary width and discrimination accuracy using R. The boundary position is the corresponding value on the x-axis of the 50% cross-over point, which can also be called as the shift point at which the most likely choice of emotion shifts from the emotion at one pole to another. The boundary width refers to the linear distance on the x-axis from the 25th percentiles to the 75th percentiles. A smaller value of boundary width represents a more rapid change in emotional perception and vice versa. For the simplicity of calculation, the 11 stimuli of each continuum were renumbered, according to the stimulus order (from left to right) in the horizontal axes of the following figures. To better evaluate the participants’ performance in discrimination task, the results were divided into two parts according to the results of identification task: (1) “peak” values, which refers to the discrimination accuracy for the emotional stimulus pairs that crossed the derived identification boundary (i.e., 50% cross-over point). (2) “nonpeak” values, i.e., the mean discrimination accuracy for the rest of the emotional stimulus pairs [2].

3. Results

3.1 Identification

The mean decoding accuracy for the three prototypes all reached 100%, indicating that no participants showed difficulty in identifying the happiness, sadness and neutral prototypes. The total identification results in two groups for each continuum are presented in Figure 1 to Figure 3.

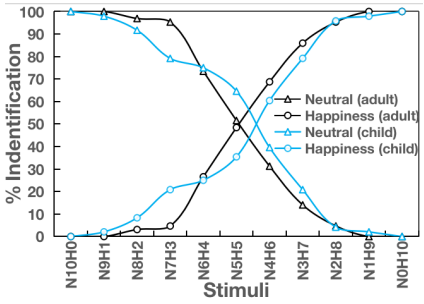

Figure 1: *Decoding accuracy of each stimulus in neutral-happiness continuum. Empty triangle, circle, square refer to identification rates of neutral, happiness, sadness, respectively. Black and blue lines refer to the adult and child group, respectively.*

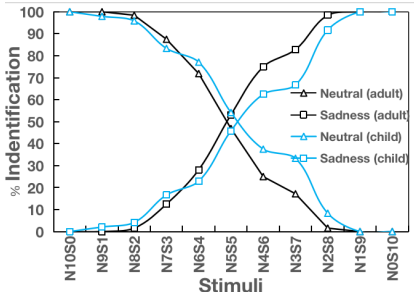

Figure 2: *Accuracy of each stimulus in neutral-sadness continuum.*

Figure 3: *Accuracy of each stimulus in happiness-sadness continuum.*

On the whole, the identification curves of the three emotional continua had the general sigmoidal shapes. Each continuum was divided into two regions. Slight raising or lowering can be found on both ends of each continuum with a sharp shift at the center of each continuum. Moreover, identification accuracy in both groups was relatively higher at

the two ends of each continuum, above 90%. The values of the identification boundary positions and boundary widths across two groups obtained by Probit analysis are shown in Table 1.

Children group. The derived boundary position was 6.24, 6.37, 5.24 in neutral-happiness, neutral-sadness, happiness-sadness continuum, respectively, where the proportion of happiness to neutral was 52.4% to 47.6%, 53.7% sadness to 46.3% neutral, and 42.4% sadness to 57.6% happiness. Children got comparative performance in identifying neutral-happiness and neutral-sadness (Table 1). Besides, the boundary position in happiness-sadness continuum was more advanced than the rest two. *Adults Group*. The derived boundary position was 6.17, 5.98, 5.44 in neutral-happiness, neutral-sadness, happiness-sadness continuum, respectively, where the proportion of happiness to neutral was 51.7% to 48.3%, 49.8% sadness to 50.2% neutral, 44.4% sadness to 55.6% happiness. The boundary position in happiness-sadness continuum was slightly more advanced than the rest two. Moreover, the boundary width in this continuum is the narrowest.

Table 1: *Derived boundary position and width of three continua. N-H, N-S, S-H refer to neutral-happiness, neutral-sadness, happiness-sadness, respectively.*

Emotional continuum	Position *M(SD)*		Width *M(SD)*	
	Children	Adults	Children	Adults
N-H	6.24(1.29)	6.17(0.88)	1.68(1.44)	1.33(1.16)
N-S	6.37(1.15)	5.98(0.97)	2.00(1.13)	1.39(1.03)
H-S	5.24(1.11)	5.44(0.91)	1.52(1.33)	0.65(0.64)

To further investigate the potential effect of age and gender on the perception of emotional continua, the repeated-measures ANOVA was conducted using R.

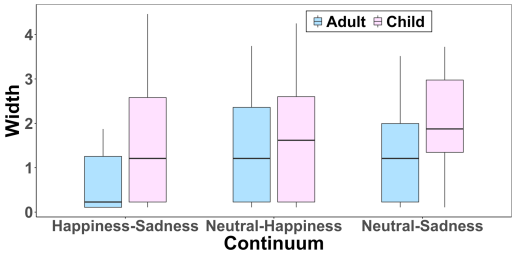

Figure 4: *Boxplot of boundary width in adult and child groups across three continua.*

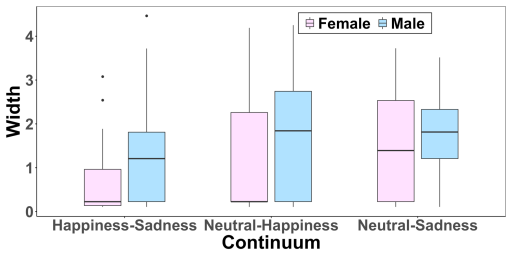

Figure 5: *Boxplot of boundary width of males and females across three continua.*

Boundary width. Mauchly's test indicated that the assumption of sphericity has been met and the results showed that boundary width was significantly affected by group [F (1,54) = 11.16, $p < 0.01$, $\eta^2 = 0.07$] and emotional type [F (2,108) = 4.41, $p < 0.05$, $\eta^2 = 0.05$]. Post hoc analysis showed that adults' boundary was significantly narrower than children's ($p < 0.001$). There was no interaction effect between group and continuum type [F (2,108) = 0.77, $p > 0.05$, $\eta^2 = 0.009$]. Additionally, there was no interaction effect between type of emotional continuum and gender. However, the results showed that gender could significantly affect boundary width [F (1,54) = 6.65, $p < 0.05$, $\eta^2 = 0.05$]. Post hoc analysis indicated that females' boundary was significantly narrower than males' ($p < 0.01$). *Boundary position*. Mauchly's test indicated that the assumption of sphericity had been violated, therefore Greenhouse-Geisser corrected tests were reported ($\varepsilon = 0.89$). The results showed that the boundary width was significantly affected by the type of continuum [F (1.78, 96.12) = 11.84, $p < 0.01$, $\eta^2 = 0.13$]. There was no significant effect of group [F (0.89, 48.06) = 0.31, $p > 0.05$, $\eta^2 = 0.002$] and gender [F (1, 54) = 0.01, $p > 0.05$, $\eta^2 = 0.00007$]. There were neither interaction effect between group and continuum type [F (0.89, 48.06) = 1.08, $p > 0.05$, $\eta^2 = 0.01$], nor interaction effect between gender and continuum type [F (2, 108) = 0.25, $p > 0.05$, $\eta^2 = 0.003$].

3.2 Discrimination

In our study, the discrimination accuracy for the stimulus pairs in which the 30% morphs coupled with the 50% morphs were the peak pairs for all the three emotional continua in two groups except for neutral-sadness continuum in the adult group, which was 50% morphs coupled with the 70% morphs.

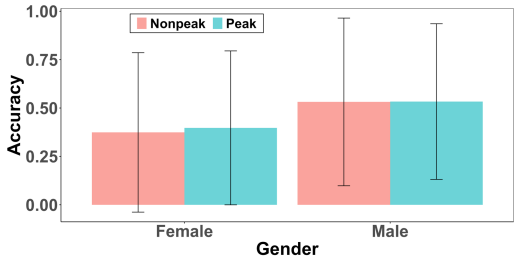

Figure 6: *Bar plot of mean discrimination accuracy of males and females, in terms of peak pairs (i.e., the stimulus pairs that crossed the derived identification boundary) and nonpeak pairs (the remaining stimulus pairs). Error bars represent 95% confidence intervals.*

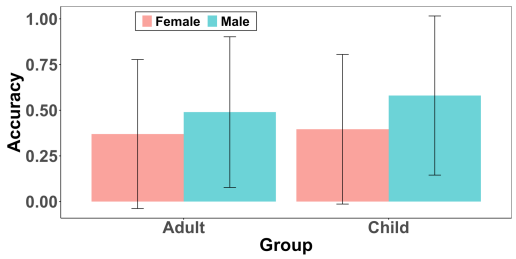

Figure 7: *Mean discrimination accuracy of males and females across two groups. Error bars represent 95% confidence intervals.*

The repeated-measures ANOVA was conducted by R. Mauchly's test indicated that the assumption of sphericity has been met and the results showed a significant effect of the type of emotional continuum [F (2,220) = 142.5, $p < 0.01$, $\eta^2 = 0.41$]. *Post hoc* analysis indicated that emotional pairs in neutral-

sadness continuum were more difficult to be discriminated than those in neutral-happiness and happiness-sadness continua (ps < 0.001). Moreover, accuracy in discriminating emotional pairs in happiness-sadness continuum was also significantly higher than those in neutral-happiness continuum ($p < 0.001$). A significant effect of gender on discrimination accuracy was also detected [F (1,54) = 11.66, $p < 0.01$, $\eta^2 = 0.09$] and *Post hoc* analysis indicated that females' discrimination accuracy was significantly lower than males' ($p < 0.01$) (Figure 6). According to the two-way ANOVA analysis, a main effect of group was also detected [F (1,666) = 7.63, $p < 0.01$] (Figure 7). *Post hoc* analysis indicated that discrimination accuracy in adult group was significantly lower than child group ($p < 0.05$). We also found that the mean discrimination accuracy of peak pairs was slightly higher than that of nonpeak pairs in general (Figure 6). However, we failed to find a significant effect of the stimuli type (peak/nonpeak) on the discrimination accuracy [F (1,110) = 0.10, $p > 0.05$, $\eta^2 = 0.0004$].

4. Discussion

This study attempted to investigate the potential categorical perception effect of vocal emotion expression in Mandarin Chinese and further detect factors that may have an effect on the perception of vocal emotion. Based on the results, we detected CP phenomena and found that age and gender were two significant factors that may play an essential role in emotional prosodic perception.

CP phenomena are explicitly reflected in identification task. The sequential emotional stimuli within each continuum were given one label on one side of a category boundary and another on the other side. Participants' identification curves showed the clear sigmoidal shape with a small slope near two endpoints and a sharp shift around the center, which is characteristic of the typical pattern of categorical perception of emotion. All these detected phenomena indicate that there is a potential CP effect during Mandarin speakers' perception of vocal emotion. This result is indeed in line with the dominant view of CP that believes categorical perception is an inborn phenomenon. Additionally, we found that as the distance from emotional prototype increased, the slope of identification curve increased. This actually implies that not all sounds expressing the same emotion can be regarded as equally good examples of that emotional category. Vocal emotions are perceptually coded in terms of their conformity to the universally accepted prototype expressions, consistent with basic emotions [2]. However, participants do not show a significant difference between discriminating peak pairs and nonpeak pairs across two groups. This result does not entail the other criteria for CP: greater sensitivity to physical change in pairs cross the categorical boundary than to the change occurring within the perceptual category [3]. This divergence may partially result from the fact that all boundary positions derived from fitting models approach to one end of the discrimination pairs. For example, in neutral-happiness continuum of adult group, the derived 50% cross-over point was 6.17, where the proportion of happiness to neutral was 51.7% to 48.3%. Therefore, we have to assume that the boundary position lies in 51.7% happiness, thus peak pair is 30% morphs (30% neutral mixed with 70% happiness) coupled with 50% morphs (50% neutral mixed with 50% happiness). We can see that the boundary is quite near to the 50% morphs. This may lead to less advantage in peak pair discrimination. Thus, no significant difference between peak pairs and nonpeak pairs, failure to correspond to the CP phenomena detected in identification tasks. However, strictly speaking, no matter how mature a certain experimental paradigm can be, there are still some limitations. We must come to the conclusion with caution since the theoretical description of CP is still under discussion.

Age plays an essential role in vocal emotion perception. Although identification curves showed explicit CP phenomena in both groups, adults have significantly narrower boundary widths than children. Their identification curves showed much shaper shift from one emotion to another in the middle of two prototypes. Interestingly, adults showed less sensitivity in distinguishing the nuance between similar stimulation. This may also indicate that adults develop their categorization ability during development and raise the possibility to classify similar sounds into one same category. On the contrary, children showed wider boundary width and relatively slower shift between two emotions in their identification curves. Their discrimination accuracy was much higher, which indicates that children are more sensitive to sound differences compared to adults, they are more likely to regard sensory differences as more meaningful and significant sensory cues and classify these differences as two different categories. Our results are also in line with the findings that age effect exists in the children's identification performance and accurate perception and interpretation ability of facial emotions continues to develop across development [23, 24]. Gender is another important factor that may have an effect on Mandarin speakers' auditory perception of vocal emotion. This study found that females showed narrower boundary width and shaper shift between two emotions in their identification curves. Similarly, they are less sensitive to auditory differences and tend to classify two similar emotional sounds into one category, which suggests that females may develop a more mature categorization sensory system compared to males.

5. Conclusions

This study is the first attempt (to our knowledge) to investigate CP effect of vocal emotion in Mandarin Chinese. According to the findings, Mandarin speakers' sensory patterns to vocal emotion expressions manifest certain CP phenomena, especially in identifying emotional morphs. However, less categorization was detected in discriminating emotional morphs. Moreover, our findings illuminate the effect of age and gender on perceptive ability of emotional speech. Adults have higher categorization ability compared to children and females also show more explicit CP phenomena than males. Our research adds to the mounting evidence on the gender difference in emotion perception. The performance divergence of the two age groups also suggests that children's ability to decode vocal emotions follows a slow course of development before turning to adults' sensory patterns. However, our study is not without its limitations as we just focused on short stimuli with a restricted view of how emotional prosody develops in more dynamic and longer utterances. More evidence is required to explore the potential CP effect in vocal emotion perception by including more emotional categories and using more experimental paradigms.

6. Acknowledgements

This research was funded by the major Program of National Social Science Foundation of China (No. 18ZDA293) and the Youth Project of Humanities and Social Sciences Foundation of the Ministry of Education in China (No. 18YJC740103).

7. References

[1] A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffit, "The discrimination of speech sounds within and across phoneme boundaries," *Journal of Experimental Psychology*, vol. 54, no. 5, pp. 358–368, 1957.

[2] P. Laukka, "Categorical perception of vocal emotion expressions," *Emotion (Washington, D.C.)*, vol. 5, no. 3, pp. 277-295, 2005.

[3] S. D. Harnad, *Categorical Perception: The Groundwork of Cognition*. Cambridge: Cambridge University Press, 1987.

[4] N. A. Macmillan, H. L. Kaplan, and C. D. Creelman, "The psychophysics of categorical perception," *Psychological Review*, 84, 452–471. vol. 84, no. 5, pp. 452−471, 1977.

[5] J. A. Bachorowski, "Vocal expression and perception of emotion," *Current Directions in Psychological Science*, vol. 8, no. 2, pp. 53-57, 1999.

[6] P. N. Juslin and P. Laukka, "Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?," *Psychological Bulletin*, vol. 129, no. 5, pp. 770-814, 2003.

[7] T. Brosch, G. Pourtois, and D. Sander, "The Perception and Categorisation of Emotional Stimuli: A Review." *Cognition and Emotion*, vol. 24, no. 3, pp. 377-400, 2010.

[8] R. Banse and R. S. Klaus, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614-636, 1996.

[9] C. E. Williams, K. N. Stevens, "Emotions and speech: some acoustical correlates," *The Journal of the Acoustical Society of America*, vol. 52, no. 4B, pp. 1238−1250, 1972.

[10] H. Levin and W. Lord, "Speech pitch frequency as an emotional state indicator," *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-5, no. 2, pp. 259−273, 1975.

[11] P. Laukka, P. Juslin, and R. Bresin, "A dimensional approach to vocal expression of emotion," *Cognition and Emotion*, vol. 19, no. 5, pp. 633-653, 2005.

[12] E. Kotsoni, M. De Haan, and M. Johnson, "Categorical Perception of Facial Expressions by 7-Month-Old Infants," *Perception*, vol. 30, no. 9, 2001, pp. 1115-1125.

[13] J. Cheal and M. Rutherford, "Categorical Perception of Emotional Facial Expressions in Preschoolers." *Journal of Experimental Child Psychology*, vol. 110, no. 3, 2011, pp. 434-443.

[14] M. Batty and M. J. Taylor, "The development of emotional face processing during childhood," *Developmental Science*, vol. 9, no. 2, pp. 207-220, 2006.

[15] J. A. Hall, "Gender effects in decoding nonverbal cues," *Psychological Bulletin*, vol. 85, no. 4, pp. 845-857, 1978.

[16] E. B. McClure, "A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents," *Psychological Bulletin*, vol. 126, no. 3, pp. 424-453, 2000.

[17] A. L. Day, and S. A. Carroll, "Using an ability-based measure of emotional intelligence to predict individual performance, group performance, and group citizenship behaviors," *Personality and Individual Differences*, vol. 36, no. 6, pp. 1443-1458, 2004.

[18] M. A. Brackett, S. E. Rivers, S. Shiffman, N. Lerner, and P. Salovey, "Relating emotional abilities to social functioning: A comparison of self-report and performance measures of emotional intelligence," *Journal of Personality and Social*, vol. 91, no. 4, pp. 780-795, 2006.

[19] M. B. Fugate, "Categorical perception for emotional faces," *Emotion Review*, vol. 5, no. 1, pp. 84-89, 2013.

[20] B. H. Repp, *Categorical perception*. New York: Academic Press, 1984.

[21] K. S. Kee, W. P. Horan, J. K. Wynn, J. Mintz, and M. F. Green, "An analysis of categorical perception of facial emotion in schizophrenia," *Schizophrenia Research*, vol.87, no.1, pp.228-237, 2006.

[22] D. J. Finney, *Probit analysis*. Cambridge: Cambridge University Press, 1971.

[23] P. M. MacDonald, S. W. Kirkpatrick, and L. A. Sullivan, "Schematic drawings of facial expressions for emotion recognition and interpretation by preschool-aged children," *Genetic, Social, and General Psychology Monographs*, vol. 122, no. 4, pp. 373-388, 1996.

[24] J. A. Russell and S. C. Widen, "A label superiority effect in children's categorization of facial expressions," *Social Development*, vol. 11, no. 1, pp. 30-52, 2002.