Impairments in Vocal Emotion Recognition in Children with Autism Spectrum Disorder: Prosody, Semantics, or Integration Difficulties? — A Three-Level Meta-Analytic Investigation

Authors: Chen Lijun, Jin Yuexin, Zeng Hanhan, Jiang Xiaoliu, Jiang Xiaoliu

Date: 2025-08-05T15:38:39+00:00

Abstract

Daily social verbal communication contains both semantic and prosodic cues. Do children with autism spectrum disorder rely on prosody or semantics to judge a speaker's emotion in social interactions? Investigating this question facilitates understanding of the disorder's etiology and provides direction for future interventions, yet it remains unresolved and vigorously debated. Accordingly, this study employed a three-level meta-analytic model to analyze 47 included studies (encompassing 93 effect sizes and 3,142 participants), conducted subgroup analyses on categorical variables (e.g., task type, cultural context, age group, control group matching type, voice gender, emotion type, and autism spectrum subtype), and performed meta-regression analyses on continuous variables (publication year, sample size, and study quality). Results revealed significant deficits in vocal emotion recognition performance among individuals with autism spectrum disorder ($g = -0.71$); integration tasks exhibited the largest effect size ($g = -0.90$), followed by prosody tasks ($g = -0.61$), with semantic tasks showing the smallest effect size ($g = -0.49$). Cultural context ($p = 0.023$) and material type within integration tasks ($p < 0.001$) moderated the vocal emotion recognition performance of children with autism spectrum disorder, and interaction effects were observed between task type and cultural context, emotion type, and autism spectrum subtype. These findings support the weak central coherence theory and provide empirical evidence for understanding the mechanisms underlying social impairment in autism spectrum disorder and for developing targeted intervention strategies.

Full Text

Preamble

Emotional Speech Recognition Deficits in Children with Autism Spectrum Disorder: Prosodic, Semantic, or Integrative Difficulties? A Three-Level Meta-Analytic Investigation

CHEN Lijun[1], JIN Yuexin[1], ZENG Hanhan[2], JIANG Xiaoliu[3]
([1] Department of Applied Psychology, School of Humanities and Social Sciences, Fuzhou University, Fuzhou 350108, China)
([2] School of Sociology and Psychology, Central University of Finance and Economics, Beijing 100098, China)
([3] Department of Social Psychology, School of Sociology, Nankai University, Tianjin 300350, China)

Abstract
Daily verbal communication contains both semantic and prosodic cues. Do children with autism spectrum disorder (ASD) rely on prosody or semantics when judging speakers' emotions in social interactions? Exploring this question is crucial for understanding the mechanisms underlying their impairments and informing future intervention directions, yet the issue remains unresolved and hotly debated. This study employed a three-level meta-analytic model to analyze 47 studies (including 93 effect sizes and 3,142 participants). Subgroup analyses were conducted for categorical variables (e.g., task type, cultural context, age group, control-group matching type, speaker gender, emotion type, ASD subtype), while meta-regression analyses examined continuous variables (publication year, sample size, and study quality). Results revealed significant deficits in emotional speech recognition among children with ASD ($g = -0.71$). The largest effect size emerged for integrative tasks ($g = -0.90$), followed by prosodic tasks ($g = -0.61$), with semantic tasks showing the smallest effect ($g = -0.49$). Cultural context ($p = 0.023$) and material type within integrative tasks ($p < 0.001$) significantly moderated performance, with significant interactions observed between task type and cultural context, emotion type, and ASD subtype. These findings support Weak Central Coherence theory and provide empirical evidence for understanding the mechanisms of social communication deficits in ASD and developing targeted interventions.

Keywords: autism spectrum disorder, emotional speech recognition, semantic cues, prosodic cues, meta-analysis

Received: April 17, 2025
This research was supported by the Fujian Provincial Education Science Planning Project (FJJKBK23-06).

Corresponding Author: JIANG Xiaoliu, E-mail: psyjxl@126.com

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by two core symptoms: deficits in social communication and interper-

sonal interaction, and restricted, repetitive patterns of behavior (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, DSM-5; American Psychiatric Association, 2015). Difficulties in emotion recognition constitute a key clinical feature of social impairment in children with ASD (Baron-Cohen, 1995), who struggle to integrate and comprehend emotions conveyed through nonverbal cues such as facial expressions or affective prosody (American Psychiatric Association, 2015).

Current research on emotion recognition deficits in ASD has predominantly focused on facial expression cues (Humphreys et al., 2007; Riby et al., 2009; Scherf et al., 2015; Li, 2016). Beyond visual channels like facial expressions and gestures (Meeren et al., 2005), speakers' voices in daily communication also contain rich emotional information. Such information may be embedded in semantic cues (Grass et al., 2016; Jaspers-Fayer et al., 2012; Mittermeier et al., 2011; Song et al., 2020), while prosodic cues such as pitch, duration, and intensity are also used to perceive different emotional states (Besson et al., 2002; Frühholz et al., 2012)—what we call "the message between the lines." Both dimensions influence speech processing (Grass et al., 2016). According to the Multi-Stage Model of Vocal Emotion Processing, emotional speech processing unfolds in three stages: first, extracting and analyzing acoustic features like pitch, loudness, and rhythm; second, tagging speech with emotional attributes; and finally, integrating multimodal cues including individual cognition, prosody, and discourse content to understand deeper socio-emotional meaning (Kotz & Paulmann, 2011; Schirmer & Kotz, 2006). Difficulties in perceiving semantic (Cicero et al., 1999) or prosodic cues (Dupuis & Pichora-Fuller, 2010), or erroneous integration of these cues during the evaluation and integration stage, can all lead to emotional speech recognition deficits.

The question of whether ASD children's emotional speech recognition deficits stem from semantic comprehension, prosodic perception, or integrative difficulties has attracted considerable attention and generated substantial controversy. First, research findings on ASD children's use of prosodic cues for emotional speech recognition are mixed. Some studies reveal that children with Pervasive Developmental Disorders (PDD; Hubbard & Trauner, 2007) show significant deficits in prosodic mastery compared to typically developing children, and that they are significantly weaker than mental-age-matched peers in recognizing happy, sad, angry, and surprised emotional prosodies (Matsuda & Yamamoto, 2015). Conversely, other studies have reached opposite conclusions: high-functioning autism (HFA) children demonstrate abilities similar to typical children in recognizing emotional intonation (Baker et al., 2010) and perceiving sentence affective prosody (Grossman et al., 2010). Second, no consensus exists regarding the use of semantic cues. Some research indicates that children with ASD screened by the Childhood Autism Rating Scale cannot effectively use semantic information to accurately judge others' emotional states (Haviland et al., 1996), while other studies show that children with Asperger Syndrome (AS) have abilities comparable to mental-age-matched individuals in recognizing emotional semantic vocabulary (Lindner & Rosén, 2006), with no significant dif-

ferences in brain activation for emotional words compared to healthy controls (Han et al., 2014). Third, presenting conflicting semantic and prosodic information simultaneously to observe which cue ASD children rely more heavily on offers another pathway to explore the origins of their deficits, yet this issue remains contentious. Some studies suggest that HFA children show excessive attention to and reliance on semantic information, neglecting prosodic cues and struggling to integrate the two (Liang et al., 2024), while other research finds that adolescents with PDD in mainstream schools, like typical adolescents, can rely more on prosodic cues for emotional speech recognition when the two cues conflict; however, PDD adolescents in special schools, similar to 6–8-year-old typical children, show no clear prosodic preference and rely more on lexical content (Segal et al., 2014). Thus, while semantic, prosodic, and integrative difficulties have each received some experimental support, no meta-analysis has compared the three. This study aims to synthesize and compare these potential causes to elucidate the underlying mechanisms.

Weak Central Coherence (WCC) theory provides a systematic framework for explaining these contradictory findings in ASD individuals' emotional speech recognition from a cognitive processing perspective. According to WCC, the core deficit in ASD children lies in their diminished capacity to dynamically integrate multimodal social cues, manifesting as excessive focus on local information and weak integration of global context (Frith & Happé, 1994). This "local advantage/global disadvantage" processing characteristic creates unique challenges for emotional recognition, which requires simultaneous parsing and integration of multidimensional information including prosody, semantics, and social context. Integration efficiency is further modulated by both sample characteristics and experimental design features (Leung et al., 2022). Regarding sample characteristics, first, cultural context shapes how social information is expressed and interpreted, thereby affecting integration demands. Second, with age, some ASD individuals may develop "rule-based" integration strategies through accumulated experience. Additionally, prior research often grouped participants by functional level or subtype (e.g., high- vs. low-functioning autism, Asperger syndrome), which may influence emotional recognition patterns. Regarding experimental design features, first, task type in emotional speech recognition—whether semantic, prosodic, or integrative—likely directly affects performance. Second, complex emotions, compared to basic emotions, depend more heavily on multichannel coordination and social-cognitive integration. Moreover, control-group matching type (e.g., IQ matching) can exclude confounding cognitive abilities and highlight the specificity of integrative deficits. Finally, speaker gender differences may affect local processing advantages through acoustic features (e.g., pitch), indirectly reflecting the rigidity of ASD children's integration strategies. Therefore, this study examines potential causes and moderating variables of ASD children's emotional speech recognition deficits from multiple perspectives, including sample characteristics (age, cultural context, ASD subtype), experimental design features (control-group matching type, emotion type, speaker gender, task type), and interactions among key variables, which

holds guiding significance for future social development and intervention for ASD children.

1.1.1 Sample Characteristics

(1) Cultural Context. Based on Hall's concept of high- and low-context cultures (Edward Hall, 2010), cultural differences may significantly influence the development of ASD children's emotional speech recognition abilities. In high-context cultures, communication tends to be more indirect, emphasizing non-verbal cues to convey information, with oral comprehension heavily dependent on context—characteristic of countries like China, Japan, and South Korea. In contrast, low-context cultures feature more direct verbal expression, using language as the primary vehicle for communication with less attention to non-verbal cues, as seen in the United States, Germany, and the United Kingdom. The social tradition of conveying deep intentions through paralinguistic cues like prosody and facial expressions in high-context cultures (e.g., East Asia) has been empirically shown to shape children's early social-cognitive patterns. For instance, Japanese preschoolers are more adept than German and American peers at using intonational cues to interpret emotions (Matsui et al., 2017), and this difference is independent of executive function and theory-of-mind levels (Ikeda et al., 2021). It can thus be inferred that ASD children growing up in high-context cultures, through long-term exposure to social scenarios reinforced by paralinguistic cues, may develop more efficient compensatory strategies for emotional speech recognition, potentially outperforming ASD groups from low-context cultures.

(2) Age. Age is a key factor moderating the degree of reliance on semantic versus prosodic cues. Due to immature executive function development, typical children struggle to shift attention from semantic to prosodic cues (Friend & Bryant, 2000), leading younger children (e.g., 6-year-olds) to have difficulty recognizing emotions through prosody even when explicitly instructed to ignore semantics (Morton et al., 2003; Waxer & Morton, 2011). Children only begin integrating prosodic cues around age 10 (Friend & Bryant, 2000). Similarly for ASD children, neuroplasticity may promote compensatory mechanisms with age, aiding basic emotion recognition (Grossman et al., 2000). Moreover, this compensatory effect may accumulate over time: Liang et al. (2024) found that 4–8-year-old HFA children relied on semantic cues under conflict conditions, whereas Brennand et al. (2011) found no significant deficits in a 10.5–19.3-year-old Asperger group. Although subtype (HFA vs. AS) and task differences may influence results, age remains a critical explanatory factor—older ASD individuals may develop cue integration strategies through long-term adaptation.

(3) ASD Subtype. Although DSM-5 unified ASD into a single spectrum, the heterogeneity within the ASD population is substantial, with intelligence and language development significantly affecting functional outcomes (Goodwin et al., 2017; Chiang et al., 2018). For example, children with Asperger Syndrome, whose language abilities approximate typical development (Gyurjyan,

2004; Spyridoula et al., 2021), may outperform other subtypes in semantic-dependent emotional speech tasks (Lindner & Rosén, 2006), yet still show deficits in prosodic sensitivity (Schelinski & von Kriegstein, 2019). Despite normal intelligence, HFA children may experience significant difficulties in complex tasks requiring semantic-prosodic integration due to weak central coherence (Uljarevic & Hamilton, 2013). Since children with Asperger Syndrome generally exhibit superior language fluency and complexity compared to HFA children, and emotional speech recognition is closely related to language ability (Zhong, 2015; Taylor et al., 2015), this may lead to better performance in emotional speech recognition tasks. Therefore, considering differences in language ability and task performance across ASD subtypes, we included it as a moderating variable.

1.1.2 Experimental Design Features

(1) Task Type. Task type (semantic, prosodic, or integrative) is a critical dimension for parsing the mechanisms of ASD children's emotional speech recognition and serves as the core entry point for exploring the origins of their deficits. Semantic tasks require decoding emotions through linguistic logic and semantic analysis (e.g., vocabulary, syntax, and context integration). Prosodic tasks depend on nonverbal auditory processing to extract acoustic features like intonation, pitch, and rhythm. Integrative tasks demand coordination of multimodal information—achieving complementary reinforcement when semantics and prosody are congruent, and requiring interference suppression (e.g., ignoring contradictory semantics to prioritize intonation) when they conflict (Leung et al., 2022; Wagener et al., 2021). Although single-dimension tasks (semantic or prosodic recognition) also involve information integration, their complexity is far lower than integrative tasks requiring cross-modal coordination. Comparing performance across the three task types can reveal the origins of ASD children's emotional speech recognition deficits: if semantic task performance is normal but integrative tasks show significant lag, this points to integrative deficits; if prosodic tasks are impaired but semantic tasks show advantages, this reflects overreliance on linguistic logic. Existing research has focused on overall task difficulty (Yeung, 2022) or speech expression forms (Zhang et al., 2022) but has not systematically distinguished task types, making it difficult to identify the root causes of deficits. This study provides an explanatory framework for the cognitive heterogeneity of ASD children's emotional speech recognition deficits by observing and comparing the magnitude of differences between ASD and typically developing children across the three task types.

(2) Control-Group Matching Type. The choice of control-group matching directly determines the comparability of study conclusions. When ASD and typical groups are unmatched, results may be exaggerated; when groups are matched on different factors, results may also differ. Research shows that chronological age (Fein et al., 1992), full-scale IQ (Jones et al., 2011), verbal IQ (Golan et al., 2007), and performance IQ (Hillier & Allinson, 2002) are

all associated with emotion recognition performance in ASD. Some studies using chronological age matching found that ASD children showed poorer ability in recognizing emotional adjectives (Van Lancker et al., 1991) and deficits in affective prosody recognition (Ozonoff et al., 1990), reflecting developmental delay. However, when matched on mental age, no significant differences emerged in affective prosody recognition (Ozonoff et al., 1990). Yet Järvinen-Pasley et al. (2008) found that even when matched on age, verbal ability, and performance IQ, ASD children still differed significantly from controls in emotional prosody recognition. Thus, whether control-group matching type influences ASD children's emotional speech recognition outcomes warrants further investigation.

(3) Speaker Gender. Male and female voices differ significantly in length, intensity, pitch, and timbre, making speaker gender a potentially important factor affecting emotional speech recognition performance. Groen et al. (2008) found no significant difference in voice gender recognition accuracy between HFA and typically developing children. Using Chinese materials, typical Chinese children showed higher accuracy in identifying male voices than female voices, whereas ASD children did not exhibit this difference (Lin et al., 2021). Beyond participant type differences, Chinese being a tonal language may also contribute to divergent results. Chinese typical children's differential recognition of voice gender may stem from male voices' pitch and other features being more easily identifiable, while ASD children's lack of significant difference in voice gender recognition suggests difficulties in processing voice gender information. Voice gender recognition relies more on pitch and timbre features, whereas emotional speech recognition requires comprehensive analysis of prosody, pitch, loudness, and other features. Although no direct evidence indicates that speaker gender affects ASD children's emotional speech recognition, their performance in voice gender recognition tasks and the shared elements between emotion and gender recognition warrant further investigation.

(4) Emotion Type. ASD children show different performance in recognizing basic versus complex emotions from speech. Basic emotions (e.g., happiness, anger, fear, sadness, surprise, disgust) show cross-cultural consistency (Ekman, 1993), whereas complex emotions are more culturally dependent, with interpretation relying on individual experience, beliefs, and cognitive states (Izard, 2007; Harris, 1991). Controversy exists regarding ASD children's recognition of basic emotions: some studies find deficits across multiple channels (facial, bodily, vocal) (Smith et al., 2010; Philip et al., 2010; Sucksmith et al., 2013), while others find no such impairments (Grossman et al., 2010). In contrast, ASD children show more consistent deficits in complex emotion recognition (Capps et al., 1992; Rosenblau et al., 2017). Complex emotion recognition requires integrating multichannel cues including semantic content, prosody, and nonverbal visual information (Fridenson-Hayo et al., 2016; Herba & Phillips, 2004), and ASD children may only utilize partial cues in recognition tasks. Therefore, this study examines the influence of emotion type on ASD children's emotional speech recognition performance, particularly whether recognition deficits exist for basic versus complex emotions.

1.2.1 Task Type × Age Group

ASD children's emotional speech recognition performance may be influenced by the interaction between task type and age group. As previously mentioned, ASD children show varying degrees of reliance on semantic and prosodic cues across different age groups (Friend & Bryant, 2000; Liang et al., 2024). Some studies found that ASD adults are more sensitive to prosodic changes than ASD children (Charpentier et al., 2018), suggesting that ASD adolescents may outperform younger children in prosodic emotion recognition tasks. Additionally, with age, ASD individuals may develop compensatory strategies for emotion recognition (Fridenson-Hayo et al., 2016), meaning that their emotion recognition deficits (compared to typical individuals) may become less pronounced. However, other research shows that emotion recognition skills in ASD patients lack improvement after late childhood, whereas typical children's skills continue maturing into adulthood (Rump et al., 2009; Uono et al., 2011). This inconsistency likely relates to task complexity across studies (Leung et al., 2022). Thus, ASD children's ability to recognize emotions across semantic, prosodic, and integrative tasks may vary by age group, indicating an interaction between task type and age.

1.2.2 Task Type × Cultural Context

ASD children face different levels of difficulty in emotion recognition across high- and low-context cultures depending on task type. Low-context communication modes center on explicit semantics, whereas high-context modes rely more on nonverbal cues to convey emotional information, requiring multimodal integration and significantly increasing recognition difficulty. At the semantic level, high-context environments often require inferring implied meaning from context, which ASD children, prone to literal interpretation, may misjudge (Segal et al., 2014). While low-context explicit emotional vocabulary simplifies recognition, high-level semantics (e.g., metaphors) remain challenging (Lampri et al., 2024). At the prosodic level, high-context cultures feature more subtle and multifunctional intonational changes (e.g., soft tones may indicate politeness or dissatisfaction), which ASD children struggle to capture. In low-context environments, prosodic emotions are more directly mapped and easier to master. At the integrative level, high-context cultures frequently present prosodic-semantic conflicts (e.g., sarcasm) requiring coordinated analysis of context, intonation, and semantics (Liu et al., 2022). In integrative tasks, ASD children's emotion recognition is poorest when emotional information is inconsistent (Stewart et al., 2013; Wagener et al., 2021). Low-context cultures, with higher information consistency, reduce integration difficulty. Notably, children in high-context cultures, through long-term exposure to nonverbal cues, may develop compensatory strategies (e.g., relying on visual aids to understand implicit emotions), though core deficits in multimodal integration persist. We therefore hypothesize that differences between high- and low-context ASD children will be most pronounced in integrative tasks, especially in complex situations requiring pro-

cessing of contradictory emotional cues.

1.2.3 Task Type × Emotion Type

ASD children's emotional speech recognition performance may be influenced by the interaction between task type and emotion type. First, ASD children exhibit unique cognitive patterns in emotion processing and relatively weak multimodal information integration abilities. Research shows that ASD individuals have weaker parsing abilities for linguistic content in semantic tasks, which may affect recognition of context-dependent emotions like surprise (Liu et al., 2022). In prosodic tasks, despite heterogeneity in sensitivity to nonverbal cues like pitch, certain emotions (e.g., anger, sadness) may be more easily identified through salient acoustic features (Juslin & Laukka, 2003). Integrative tasks require coordinating semantic and prosodic information, and ASD's weak central coherence tendency may cause difficulties in recognizing complex emotions like sarcasm or contradictory feelings (Uljarevic & Hamilton, 2013). Second, emotion type may further modulate task performance differences due to varying degrees of reliance on contextual and sociocultural cues. Basic emotions have clear adaptive survival value with prominent biological features, whereas complex emotions are more shaped by sociocultural factors, requiring integration of social contexts, and some even contain mixed valences (e.g., bittersweet feelings) (Izard, 2007). When recognizing complex emotions in integrative tasks requiring both prosodic and semantic coordination, ASD children may show fluctuations due to cognitive overload, performing particularly poorly with high cognitive-load emotions. Therefore, the ASD population's unique emotion processing characteristics and the inherent features of emotion types may lead to varied recognition performance across tasks for basic versus complex emotions.

1.2.4 Task Type × ASD Subtype

ASD children show significant heterogeneity in emotional speech recognition task performance, which is not only related to the cognitive demands of task types but may also be deeply modulated by the neurodevelopmental characteristics and cognitive processing patterns of ASD subtypes. First, language abilities across subtypes may drive task performance dissociation: Asperger Syndrome (AS) children have relatively intact semantic integration functions, enabling better performance in semantic-dependent tasks than other subtypes (Lindner & Rosén, 2006). Second, in prosodic tasks, both HFA and AS children show attention deficits due to abnormal activation of the salience network causing imbalanced attention resource allocation (De Giambattista et al., 2019). They may overfocus on acoustic details, showing selective sensitivity to high-intensity emotions (e.g., anger) in prosodic tasks, yet remain deficient in recognizing low-intensity or complex prosodies (e.g., sarcastic rising tones) (Järvinen-Pasley et al., 2008), meaning the gap between the two subtypes may not be obvious in prosodic emotion recognition tasks. Third, from a cognitive processing perspective, AS and HFA share relatively good cognitive and language abilities with

normal intelligence (Koyama et al., 2007), yet neuroscientific evidence shows they exhibit insufficient amygdala and orbitofrontal cortex activation when processing emotional information, with impaired ability to extract and integrate emotional information, manifesting as relatively low integration in higher-order brain regions and overly strong sensory area functions (Fan et al., 2020). Therefore, they may experience systematic failures in tasks requiring coordinated semantic and prosodic analysis (Uljarevic & Hamilton, 2013). Consequently, all subtypes may perform poorly in integrative tasks due to central coherence deficits, meaning the gap between AS/HFA individuals and other subtypes (especially low-functioning ASD) may be less pronounced in integrative than in semantic tasks. However, existing research has largely examined main effects of task type or subtype classification in isolation, lacking systematic integration of their interaction mechanisms.

In summary, research findings remain inconsistent and inconclusive regarding whether children with autism spectrum disorder show clear deficits in emotional speech recognition, and if so, whether these originate from semantic comprehension, prosodic exploration, or difficulties integrating the two. This study will integrate relevant domestic and international literature to examine the influence of semantic and prosodic cues on ASD children's emotional speech recognition, clarifying whether deficits exist and their underlying causes. Additionally, the study will explore moderating effects of publication year, sample characteristics (cultural context, age, ASD subtype), experimental design features (task type, speaker gender, emotion type, control-group matching type), sample size, and study quality on ASD children's emotional speech recognition performance. By deeply investigating these difficulties, this study not only helps understand challenges in social communication but also informs the design of more targeted intervention programs to help them better integrate into society and engage in social activities, while providing evidence for developing comprehensive emotion recognition intervention strategies.

2 Methods

This study was preregistered on the Open Science Framework platform (DOI: https://doi.org/10.17605/OSF.IO/5P3MC).

2.1 Literature Search

The search procedure was as follows: Using the keyword combination ((“ASD” OR “autism spectrum disorder” OR “autism*” OR ”ASC” OR ”Asperger” OR ”HFA”) AND (”affect*” OR “emotion*”) AND (”perception*” OR “recognition” OR “process*” OR ”comprehension*” OR “interpretation” OR “judge*” OR ”rate*” OR “decode*”) AND (”prosody*” OR “lexical*” OR ”semantics*” OR “word” OR “pragmatic*” OR ”grammar” OR ”language” OR ”speech” OR ”content” OR ”verbal” OR ”vocal*”)), we searched three English databases: Web of Science, PubMed, and ProQuest. Using the Chinese keyword combination ((“自闭症” OR “孤独症” OR “阿斯伯格”) AND (情绪识别 OR 言语理解) AND (韵律 OR 语

义 OR 词义 OR 句义 OR 语音 OR 口语)), we searched three Chinese databases: CNKI, VIP, and Wanfang. The search period was set from January 1, 1980, to December 28, 2024, as autism research shifted toward cognitive and social deficits in the 1980s, making emotion recognition a key research focus. To ensure comprehensive retrieval, Google Scholar was used for supplementary searches, and a snowballing method was applied to examine valuable references cited in retrieved articles, along with manual searches of reviews and related articles, yielding 1,548 articles.

2.2 Inclusion and Exclusion Criteria

Inclusion criteria: (1) Participants must be children with autism spectrum disorder, with explicit age variables (under 18 years) or "child" designation; (2) Studies must include at least two groups: an experimental group (ASD children) and a typically developing control group; (3) Studies must include emotional speech recognition tasks with complete data convertible to effect sizes; (4) For duplicate publications, only one was selected.

Exclusion criteria: (1) Book chapters, review articles without original data, conference proceedings, and patents; (2) Studies with participants all over 18 years or other special populations; (3) Participants with hearing impairments or comorbid affective disorders like depression; (4) Studies on populations unrelated to ASD. The literature screening process is shown in Figure 1 [Figure 1: see original paper].

2.3 Data Extraction and Quality Assessment

Figure 1. Literature Search and Screening Flowchart

Eligible studies were coded for the following variables: (1) Author surname and publication year; (2) Participant location: country/region and cultural classification. According to Hall's classification criteria (Zhao & Zeng, 2009), high-context cultures include Asia, Latin America, and Africa, while North America and Northern Europe are considered low-context cultures; (3) Age information for ASD and control groups: chronological age and age group classification based on Erikson's psychosocial development theory (Erikson, 1994), dividing participants into preschoolers (3–6 years), children (6–12 years), and adolescents (12–18 years). Specific rules: (a) age ranges completely within a single segment were labeled accordingly; (b) ranges clearly spanning two or more age groups that could not be separated were labeled "mixed" and excluded from age-related subgroup and interaction analyses; (c) ranges slightly deviating from preset boundaries (by about 1 year), or where mean age could be assigned to a specific group and the original article used clear labels like "preschooler/young child," "school-age child," or "adolescent," were assigned to the corresponding single age group to maximize information preservation and mitigate statistical power issues from category imbalance; (4) Gender ratios for ASD and control groups; (5) Sample sizes; (6) Emotional speech recognition task type: semantic,

prosodic, or integrative. Semantic tasks require identifying emotions conveyed by semantics; prosodic tasks require identifying emotions conveyed by prosody; integrative tasks require identifying emotions conveyed by speech overall (without specifying prosody or semantics); (7) Control-group matching type: whether control and ASD groups were matched on mental age, chronological age, verbal IQ, performance IQ, or full-scale IQ—if any matching was successful, it was recorded; (8) Speaker gender: male only, female only, or both; (9) Emotion type: basic emotions include happiness, anger, fear, sadness, surprise, and disgust (Ekman, 1982), with all others classified as complex emotions; (10) Material type: (i) prosody-expressing materials: neutral or non-semantic content (e.g., nonsense syllables) with emotion conveyed only through prosody; (ii) semantics-expressing materials: neutral prosody with emotion conveyed only through semantics; (iii) congruent materials: both prosody and semantics convey clear, identical emotional information; (iv) incongruent materials: prosody and semantics convey clear, contradictory emotional information; (v) combined materials: containing both semantic and prosodic emotional cues (e.g., comprehensive videos) without clear congruence or conflict, requiring judgment based on combined cues.

Data extraction followed the principle of using independent samples as the unit for effect size extraction; when a single article provided multiple effect sizes, each was coded separately. Referencing previous literature quality assessment systems (Ren et al., 2023; Zhang et al., 2019), this study evaluated included literature using the following criteria: (1) Control-group matching (0–2 points): 0 = no matching or unsuccessful matching on any of mental age, chronological age, verbal IQ, performance IQ, or full-scale IQ; 1 = successful matching on any indicator but without reported specific values; 2 = successful matching with reported between-group comparison statistics; (2) Diagnostic criteria (0–2 points): 0 = no diagnostic criteria reported; 1 = diagnostic criteria reported only (e.g., ICD, DSM, clinician assessment); 2 = diagnostic criteria plus specific symptom/scale scores for the ASD group; (3) Publication type (0–1 point): 0 = unpublished dissertation; 1 = published journal article; (4) ASD sample size (0–1 point): 0 = ASD sample < 30; 1 = ASD sample ≥ 30. The sum of these scores constituted the literature quality assessment total, with higher scores indicating better quality.

2.4 Model Selection

In most original studies included in this meta-analysis, the same research sample typically involved multiple effect sizes (e.g., simultaneously assessing prosodic, semantic, and integrative tasks), creating interdependence among these effect sizes. Additionally, some studies reported multiple effect sizes from different samples, further increasing data complexity. Under these conditions, using fixed-effect or random-effects models that assume independent effect sizes could lead to biased or overestimated meta-analytic results (Lipsey & Wilson, 2001). To address these issues, this study employed a three-level meta-analytic model,

introducing an additional level to capture differences between different samples within the same study. The three-level model further decomposes variance sources into: Level 1 reflects sampling error (variance due to sampling in original studies); Level 2 captures differences between effect sizes within the same study—if heterogeneity at this level is significant, it indicates substantial differences between effect sizes within the same study; Level 3 reflects differences between effect sizes across studies—if significant, it indicates heterogeneity between studies (Cheung, 2014). Compared to traditional meta-analysis, three-level meta-analysis offers significant advantages in handling dependence of effect sizes from the same study and heterogeneity within and between studies, maximizing data retention while improving statistical power (Assink & Wibbelink, 2016). Based on this, this study utilized three-level meta-analytic models for main effect testing, heterogeneity testing, moderator analysis, publication bias assessment, and sensitivity analysis to comprehensively reveal underlying patterns and moderating effects in the data structure. Additionally, nested models were used to further control for the complexity of the multilevel data structure and more accurately capture true sources of effect size variation.

2.5 Effect Size Calculation

This study conducted meta-analyses using the metafor package in R, encompassing publication bias testing, main effect testing, and moderator effect testing. Following Nakagawa et al. (2023), the orchaRd 2.0 package was used to generate orchard plots for visualizing statistically significant moderating and interaction effects. Given that emotional speech recognition performance research typically uses means and standard deviations as evaluation criteria, this study selected the standardized mean difference Hedge's g as the effect size indicator. Hedge's g corrects for small sample sizes (Hedges, 1981) and provides more precise estimates than Cohen's d (Caron et al., 1988). For effect size magnitude, Hedge's g values of 0.20, 0.50, and 0.80 correspond to small, medium, and large effects, respectively. Notably, negative g values indicate that ASD children performed worse than typically developing controls on emotional speech recognition. To ensure robustness, this study used Cook's distance for influence analysis of effect sizes and conducted sensitivity analyses.

2.6 Publication Bias Assessment

Publication bias refers to the greater likelihood of significant results being published (Rothstein et al., 2005), leading to a higher proportion of statistically significant results in the published literature than non-significant ones. Researchers may only collect published significant literature, thereby affecting meta-analytic results. This study used funnel plots and the fail-safe number (Nfs) method to assess publication bias risk (Khoury et al., 2013).

3 Results

3.1 Literature Inclusion

This study included 47 articles comprising 93 effect sizes and 3,142 participants, including 5 dissertations and 42 journal articles, as shown in Table 1 . The quality scores of included literature ranged from 3 to 6, with a mean of 4.45 ± 0.65.

Table 1. Information on Original Studies Included in the Analysis

Author (Year)	Location	Cultural Context	Age Range (Mean) [ASD] [Control] [Years]	Gender Ratio (M/F) [ASD] [Control]	Control-Group Matching	Task Type	Material Type	Effect Size
Angeleri (2016)	Italy	Low	4.9–15.2 (9.46)	4.8–15.1 (8.73)	10/14	Chronological age, full-scale IQ	Cognitive	Semantics-expressing
Baker (2010)	-	Low	10–14 (12.8)	-	9.58$±1.0\|10–14(12.2)\|6.33±$	10.5–20.3 (14.5)	11–16.7 (13.3)	5.0–17.6 (9.0)
Chiew (2017)	-	-	-	-	-	-	-	-

Note: Figure translations are in progress. See original paper for figures.