frontiers | Frontiers in Psychology

REVIEW
published: 03 June 2022
doi: 10.3389/fpsyg.2022.906848

# How Tone, Intonation and Emotion Shape the Development of Infants' Fundamental Frequency Perception

*Liquan Liu[1,2,3]*, Antonia Götz[1,4], Pernelle Lorette[5] and Michael D. Tyler[1,3]*

[1]*MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Penrith, NSW, Australia,* [2]*Center for Multilingualism in Society Across the Lifespan, University of Oslo, Oslo, Norway,* [3]*Australian Research Council Centre of Excellence for the Dynamics of Language, Canberra, ACT, Australia,* [4]*Department of Linguistics, University of Potsdam, Potsdam, Germany,* [5]*Department of English Linguistics, University of Mannheim, Mannheim, Germany*

Fundamental frequency ($f_0$), perceived as pitch, is the first and arguably most salient auditory component humans are exposed to since the beginning of life. It carries multiple linguistic (e.g., word meaning) and paralinguistic (e.g., speakers' emotion) functions in speech and communication. The mappings between these functions and $f_0$ features vary within a language and differ cross-linguistically. For instance, a rising pitch can be perceived as a question in English but a lexical tone in Mandarin. Such variations mean that infants must learn the specific mappings based on their respective linguistic and social environments. To date, canonical theoretical frameworks and most empirical studies do not view or consider the multi-functionality of $f_0$, but typically focus on individual functions. More importantly, despite the eventual mastery of $f_0$ in communication, it is unclear how infants learn to decompose and recognize these overlapping functions carried by $f_0$. In this paper, we review the symbioses and synergies of the lexical, intonational, and emotional functions that can be carried by $f_0$ and are being acquired throughout infancy. On the basis of our review, we put forward the Learnability Hypothesis that infants decompose and acquire multiple $f_0$ functions through native/environmental experiences. Under this hypothesis, we propose representative cases such as the synergy scenario, where infants use visual cues to disambiguate and decompose the different $f_0$ functions. Further, viable ways to test the scenarios derived from this hypothesis are suggested across auditory and visual modalities. Discovering how infants learn to master the diverse functions carried by $f_0$ can increase our understanding of linguistic systems, auditory processing and communication functions.

**Keywords: lexical tone, intonation, Prosody, phonological theory, sensory processing, cognitive processing, cross-linguistic transfer, emotional tone**

OPEN ACCESS

***Edited by:***
*Hatice Zora,*
*Max Planck Institute for Psycholinguistics, Netherlands*

***Reviewed by:***
*Aijun Li,*
*Institute of Linguistics (CASS), China*
*Linda Polka,*
*McGill University, Canada*

****Correspondence:***
*Liquan Liu*
*l.liu@westernsydney.edu.au*

***Specialty section:***
*This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology*

***Received:*** *29 March 2022*
***Accepted:*** *10 May 2022*
***Published:*** *03 June 2022*

***Citation:***
*Liu L, Götz A, Lorette P and Tyler MD (2022) How Tone, Intonation and Emotion Shape the Development of Infants' Fundamental Frequency Perception. Front. Psychol. 13:906848. doi: 10.3389/fpsyg.2022.906848*

## INTRODUCTION

From the beginning of life, humans are exposed to the fundamental frequency ($f_0$; Titze et al., 2015). The $f_0$ carries a wide range of information. This includes linguistic (e.g., lexical tone), paralinguistic (e.g., speaker intent, emotion, Crystal and Quirk, 1964; Gussenhoven, 2002), and extralinguistic information (e.g., melody, Johnson, 1990; He et al., 2007). While some

crucial communicative functions carried by $f_0$ appear to be universal, such as intonation (Best, 2019), others can vary across the world's languages (e.g., signalling grammatical information; Hyman, 2011, 2016; Remijsen, 2016). For example, a syllable /ja/ with a rising $f_0$ can be recognised as an attention getter for a Dutch speaker, but as the word "tooth" for a speaker of Mandarin. Thus, to acquire the language of their environment, infants are faced with a complex task. They must learn to disambiguate, decompose, recognise, and learn the patterns of $f_0$ variability that apply to different linguistic, paralinguistic, and non-linguistic domains.

It is impressive that infants process different sources of speech information and eventually learn to disentangle functions of $f_0$ during speech perception, yet *how* they achieve this has received little attention in the empirical or theoretical literature. Research on infants' perception, production, and learning of the functions carried on $f_0$ has focused mainly on a single specific domain of interest, for example, music, lexical tone, or intonation. To explain how infants learn to perceive the multifaceted and cross-domain $f_0$ signal, it will be necessary to integrate findings across those different domains of interest. The purpose of this paper is to sketch out an approach to doing that across three $f_0$ functions: tone, intonation and emotion. We first review empirical studies on infants' acquisition of the three functions of interest along with their interactions. After that, theoretical considerations are discussed, followed by the proposal of a novel hypothesis.

## INFANTS' ACQUISITION OF TONE, INTONATION, AND EMOTION CARRIED ON $f_0$

### Tone

Around 60–70% of world languages are tonal (Yip, 2002), predominantly using contrastive $f_0$ variations to differentiate lexical and grammatical changes. Spreading across Asia, Africa, (indigenous) America, Europe and South Pacific regions (Maddieson, 2013), tone languages are spoken by more than half of the world's population (Fromkin, 2014). Among tones, the predominant $f_0$ changes lie in pitch height (level, register) and pitch direction (contour, slope; Chao, 1947; Gandour, 1983; Gussenhoven, 2004). Of particular interest are tone languages that rely on lexical tones to distinguish word meanings. For instance, the syllable [ji] in Cantonese means "cure" when bearing a high level tone, but "son" with a low falling tone (Francis et al., 2008). In a tone language such as Mandarin, $f_0$ carries the primary cues for perception (Gandour, 1983; Massaro et al., 1985; Lee and Lee, 2010), in addition to secondary cues such as intensity and duration (Jongman et al., 2006). The stark difference in $f_0$ functions in lexical-tone versus non-tone languages raises important questions about how these typological differences influence the development of speech perception, speech production, and word learning.

Speech perception research has shown clear differences in the way that speech is perceived by tone and non-tone language learning infants (Fikkert et al., 2020) as well as by adults (Burnham and Singh, 2018; Liu et al., 2022). Such studies have demonstrated increased tonal sensitivity over the first year after birth for tone language learners and decreased sensitivity for non-tone language learners (Mattock and Burnham, 2006; Mattock et al., 2008; Yeung et al., 2013). However, empirical evidence in the last decade appears to challenge these canonical patterns. For instance, there appears to be an age-based increase in sensitivity to certain tonal contrasts for both tone and non-tone language learning infants (Chen and Kager, 2016; Chen et al., 2017; Tsao, 2017; Ramachers et al., 2018; Singh et al., 2018), and behavioural and neural studies report that bilingual infants tend to be more resilient in perceiving and learning tones even when they do not exist in these infants' linguistic repertoires (Graf Estes and Hay, 2015; Liu and Kager, 2017a; Liu et al., 2019). Further, a U-shaped sensitivity has been reported in non-tone language learning infants, such that the decline in sensitivity observed over the first year of life is reversed in their second year (Liu and Kager, 2014, 2017a; Götz et al., 2018). Thus, while initial investigations into infant speech perception showed expected declines in sensitivity for tonal contrasts for infants learning a non-tone language, more recent studies suggest that the developmental trajectory requires a more nuanced theoretical interpretation (for similar observations on the development of consonant perception, see Tyler et al., 2014; Liu and Kager, 2015).

Tone production studies typically involve tone language-learning infants, who start producing $f_0$ contours around 7 months (Chen and Kent, 2009). It is unclear whether the $f_0$ produced is on a lexical or utterance level (or both), however, because adults cannot identify the ambient language when listening to the babbling of 8–12-month-old English and Mandarin-learning infants extracted from recordings (Lee et al., 2017). Mature production can be observed shortly after 2 years of age (Li and Thompson, 1977; So and Dodd, 1995; Hua and Dodd, 2000; Hua, 2002; To et al., 2013, for a review, see Peng and Chen, 2020). Recent acoustic analyses challenged this conclusion, however, as they have revealed substantial differences between children and adults' tone production. Mandarin-learning children have been found not to reach an adult level of tonetic realisation until the age of 5 (Wong et al., 2005; Wong, 2012a,b, 2013), possibly due to complex tone articulation (Wong, 2012a) or tonal rules (Chen et al., 2015; Wewalaarachchi and Singh, 2016).

The conflicting findings also extend to word learning. To learn a tone language, children need to associate lexical items that differ minimally in tonal contrasts with different word meanings. Making such associations does not appear to be easy for children at 2–3 years (Shi et al., 2017) and the lexical encoding does not stabilise until around 4–5 years (Singh et al., 2015). Sensitivity to tonal contrast is not required for non-tone language learning infants, yet they are sensitive to $f_0$ variations on words at 7.5 and 18 months (Singh et al., 2008, 2014). While 14-month-olds are able to associate non-native tones with different objects, that ability decreases at 18 months (Hay et al., 2015; Liu and Kager, 2018). By

2.5 years, they no longer consider $f_0$ change to be lexically relevant (Quam and Swingley, 2010).

Mixed findings in perception, production, and learning trajectories among tone and non-tone language-learning infants require further investigation. In this paper, we raise the hypothesis that these discrepancies can be attributed to other functional uses of $f_0$, which are linguistically and paralinguistically relevant in all spoken languages as they can also manifest on the utterance level, such as intonation.

## Intonation

All spoken languages employ intonation (Best, 2019), where $f_0$ acts at a phrasal level (distinct from the word-level tone, and in addition to other cues such as voice quality; Ladd et al., 1985). When learning a language like English, children need to know that different $f_0$ contours applied to the same utterance can signal different (e.g., narrative, interrogative) connotations. Intonation can convey linguistic information, facilitate the acquisition of other linguistic components (e.g., words; Thiessen et al., 2005), raise attention (Sullivan and Horowitz, 1983), and carry speakers' intentions (Gussenhoven, 2002; Esteve-Gibert et al., 2017). Adult listeners can encode both focus and interrogative meaning in intonation (Liu and Xu, 2005). Arguably, this makes intonation a unique component, as it spreads across linguistic and paralinguistic fields and serves grammatical, pragmatic and affective functions (Snow and Balog, 2002). Furthermore, intonation plays a crucial role in caretaker-infant interactions and communications (Stern et al., 1982; Fernald and Simon, 1984; Fernald, 1989).

Infants' perception and production of intonation develop concurrently with tone throughout infancy and early childhood. Newborns are sensitive to intonation in speech (Nazzi et al., 1998; Sambeth et al., 2008) and 6-month-olds can use pitch contours to parse utterances into clauses (e.g., Seidl, 2007). By 6 and 9 months, European Portuguese-learning infants can discriminate single prosodic-word utterances differing in statement (falling) or yes–no question (falling-rising) intonation (Frota et al., 2014; Frota and Butler, 2018). Despite their sensitivity to the $f_0$ differences that characterise intonation, children do not appear to rely strongly on intonation to signal conversational turn taking until 3 years and onwards (Keitel et al., 2013). Why they are reluctant to do so at earlier ages needs to be understood.

Arguably, intonation production starts from birth with crying (Mampe et al., 2009) and vocalisation shortly after birth (Kent and Murray, 1982). Newborn infants' crying patterns already reflect the intonation patterns of their native language (Mampe et al., 2009; Wermke et al., 2016, 2017; Manfredi et al., 2019; Prochnow et al., 2019). Infants begin with a predominant falling pitch contour then progress to other $f_0$ patterns, with accent range increasing with age (Snow, 2001). The production of pitch register stabilises in the single-word period, and core features are controlled in the two-word stage (Snow and Balog, 2002). However, the development of intonation production in the first 2 years of life is not linear. At the end of the first year after birth, rising and falling contours are produced with a smaller accent range in comparison to the 6–9 and above 18-month-olds. This U-shaped pattern needs further investigation and explanation.

With respect to the interaction between tone and intonation, researchers are prone to argue for a linguistic status of tone and an ambiguous status of intonation: from a categorical perspective, studies favour evidence for discrete tone but not intonation categories, as one "intoneme" may consist of various intonational elements (Tonkova-Yampol'skaya, 1969; but see So and Best, 2014 on "i-category"). Tone-language speakers show distinct tone and intonation processing differences on single-syllable units, not only in the neural organisation of subcortical and cortical structures but also hemispheric lateralisations (Chien et al., 2020), although to date, no consensus has been reached on whether intonation is dominantly processed in the left or right hemisphere. An utterance-final rising $f_0$ tends to be a universal cue for the perception of interrogation (Gussenhoven and Chen, 2000; Liang and Heuven, 2007), but perception of intonation appears to be tone-dependent. In Mandarin, a yes/no question is more easily identified when the utterance ends with a falling than a rising tone (Yuan, 2011), and a declarative versus interrogative contrast elicits strong mismatch negativity responses on syllables with falling but not rising tones (Ren et al., 2013). Research connecting intonation with word learning is relatively scarce. Although English speakers demonstrate the presence of long-term memory traces for prosodic information in the brain (Zora et al., 2015), English-learning 2-year-olds do not interpret salient pitch contour differences (rising-falling vs. falling-rising) as inherent to novel words (Quam and Swingley, 2010).

Such tone-intonation interaction in perception is not restricted to speakers of a tone language. Among non-tone language speakers, the component that stabilises the earliest, pitch register (Snow and Balog, 2002), can facilitate the perception of non-native tone contrasts (Liu et al., 2022). Non-tone language speakers' knowledge of intonation also appears to influence tone perception. For instance, the rising versus falling tones in Mandarin Chinese are similar to the declarative versus interrogative $f_0$ patterns in languages such as English (Braun and Johnson, 2011; So and Best, 2011, 2014). Indeed, when examining American English-learning infants' Mandarin tone-object association at 14 months, infants were more successful for words with a rising tone than for words with the other three Mandarin tones (Hay et al., 2015, 2019). This suggests that they may have been able to capitalise on their developing sensitivity to English rising pitch intonation for perception of non-native words differing by lexical tone. Adopting intonation patterns from a non-tonal native language for perception of non-native tones is consistent with theories of perceptual assimilation (Best, 1994, 2019; Best et al., 2009; So and Best, 2010, 2014), which may provide a potential theoretical explanation for the U-shaped developmental pattern reported in infant perception of non-native tones (Liu and Kager, 2014, 2017a; Götz et al., 2018). Children learning non-tone languages may become less sensitive to certain $f_0$ patterns as they recognise that tonal variations do not signal lexical distinctions in their native language, while also learning the complementary functions that *are* carried on $f_0$.

The nonlinear developmental trajectory for intonation from infancy to toddlerhood (Snow and Balog, 2002), the restricted use of $f_0$ as a cue in intonation in early childhood (Keitel et al., 2013), and the overlap between tone and intonation in adulthood across the world's languages (e.g., Gussenhoven and Chen, 2000) all highlight the need to comprehensively understand $f_0$ functions along the developmental trajectory. Additionally, research on infants' acquisition of intonation may benefit from considering the prosodic and information structures of intonation, but few studies have taken this approach (Frota and Butler, 2018). For example, according to Autosegmental-Metrical accounts (Pierrehumbert, 1980; Grice et al., 2006; Ladd, 2008; Arvaniti and Fletcher, 2020), intonation is composed of a series of tonal events. To reveal the trajectory and mechanisms infants use to recognise word- and phrase-level prosody from continuous speech, it may be necessary to take the componential structure of intonation into consideration. Further, the visual aspect of intonation, often discussed in sign languages (e.g., Dachkovsky and Sandler, 2009), it still poorly understood in spoken languages. While expressing uncertainty, speakers not only use prosodic cues such as rising intonation, but also facial cues involving eyebrow raising, head tilting, furrowing, etc. (Dijkstra et al., 2006; Roseano et al., 2016).

In the next section, we attempt to explore the $f_0$ function in the domain of emotion, as well as the entanglement between the intonational and emotional functions in speech directed to infants.

## Emotion

At first glance, there are differences in how theories consider $f_0$ between linguistic and emotional domains. This is not surprising since emotion theories typically focus on visual emotional signals (e.g., facial expressions) rather than how emotion is coded in speech. Theoretical debates centre on whether humans possess innate basic emotion categories, in both facial expressions (Chong et al., 2003; Gendron et al., 2018) and emotions in vocalisations (Sauter et al., 2010, 2015; Gendron et al., 2014, 2015). Empirical evidence suggests distinct processing of $f_0$ functions in intonation and in emotion. Emotional voice cues are processed predominantly in the auditory cortical areas in the right hemisphere, whereas phonemic cues are processed mainly in the left (Kotz et al., 2006; Scott and McGettigan, 2013). Limited studies have discussed the interaction between linguistic and emotional $f_0$ functions (Kotz and Paulmann, 2007; Pell and Kotz, 2011). It is unclear whether certain regions are responsible for $f_0$ variations in both emotional and linguistic states (Frühholz et al., 2012; Liebenthal et al., 2016).

For preverbal infants, perception of emotion is critical for survival in a social world, as it constitutes one of the critical social cognition skills. While emotion signals in the visual domain are most representative in a speaker's face and body language, they are carried primarily by $f_0$ in speech (Remez et al., 1981; Ladd et al., 1985; Scherer, 1986, 2003; Goldbeck et al., 1988). There are also secondary cues for emotion in speech (Murray and Arnott, 1993; Banse and Scherer, 1996; Bänziger et al., 2015; Pell et al., 2015), including intensity and speech rate (Scherer, 1986), pausing structure (Cahn, 1990) and duration (Mozziconacci, 1998), and timbre/voice quality (Gobl et al., 2002; Gobl and Chasaide, 2003; Yanushevskaya et al., 2018). In particular, $f_0$ modulates and strengthens the affective and motivational contexts in both infants (Stern et al., 1982) and adults (Frick, 1985). It also has an advantage over other cues, such as timbre, that it is simple to measure and quantify.

With respect to emotion perception, infants' ability to experience and perceive emotion has been hypothesised to develop as a function of neural development, increasing the capacity of processing emotional concepts with the aim of assigning meaning to sensory inputs and guiding behaviour (Hoemann et al., 2019). In their first year of life, infants are sensitive to emotions expressed from different cultures (Liu et al., 2021), and employ different attentional strategies based on their native culture (Geangu et al., 2016). Although emotional $f_0$ is highly salient in the environment from the beginning of life (ManyBabies Consortium, 2020), and its development is likely linked with the neuro-cognitive development of socio-emotions, the detailed trajectory of emotional $f_0$ remains unclear.

There appear to be $f_0$ patterns with distinct acoustic characteristics for different emotions (Liu and Pell, 2012; Wang and Lee, 2015), although findings are mixed on whether emotional $f_0$ patterns are universal or culturally-specific (Murray and Arnott, 1993; Pell et al., 2009; Li, 2015). Some perception studies have suggested a universal association between high $f_0$ and positive emotion (e.g., happiness, Ortony et al., 1990; Ilie and Thompson, 2006; Belyk and Brown, 2014), but the same trend has not been observed in other corpus studies (Laukka et al., 2005; Goudbeek and Scherer, 2010). The $f_0$ acoustics of the same emotional tone can vary across studies in height and range (Pell et al., 2009), along with other cues such as intensity and duration (Wang and Lee, 2015; Wang and Qian, 2018). Furthermore, cross-linguistic and cultural differences have been reported in both the acoustic manifestation (Douglas-Cowie et al., 2003; Anolli et al., 2008; Wang et al., 2018) and the interpretation (Koeda et al., 2013) of $f_0$. Despite this substantial variation, infants appear to identify regularities to build their knowledge.

There has been a debate in the literature on the processing of emotions in (visual) facial expressions about whether universal categories of basic emotional categories (e.g., happiness, anger) exist (Gendron et al., 2018). Infants can disambiguate between some emotional categories (Caron et al., 1985; Haviland and Lelwica, 1987; Soken and Pick, 1999; for a review, see Widen, 2013), yet it is unclear whether they conceptualise and abstract emotional features such as valence or arousal (Ruba et al., 2020). In comparison, research on processing of (auditory) vocal expressions of emotion is relatively scarce. Unlike 3-month-olds, infants at 5 months can discriminate between vocal expressions of positive and negative valence, but they do so reliably only in the presence of a face (Walker-Andrews and Grolnick, 1983; Walker-Andrews and Lennon, 1991). Infants aged 7 months process emotions of positive and negative valence differently, not only in facial expressions (Nelson and De Haan, 1996) but also in emotional prosody (Grossmann et al., 2005). With respect to the production of emotional $f_0$, a

parental rating study has shown that vocalisations of 2-month-olds can be judged to fit along a comfort-discomfort dimension (Papoušek, 1989). Infants often use prosody, including (high) $f_0$, to signal what is perceived by their caretakers as emotional cues, be it wailing of fear or crying for attention (for a review, see Bryant, 2021). Little is known about the two-way relationship of $f_0$ functions in tone and emotion, although language background (tone vs. non-tone languages) has been shown to play a role. Larger $f_0$ variations of emotional tones are produced by non-tone than tone language speakers (Ross et al., 1986; Anolli et al., 2008; Wang et al., 2018), suggesting that the lexical function of $f_0$ constrains its use for emotional function.

Some studies have shown that emotional $f_0$ can facilitate word learning (for a review, see Doan, 2010). For example, words with emotional variations are better recognised in fluent speech by English-learning 7-8-month-olds than words without such variability (Singh, 2008). Infants aged 10.5 months showed significant positive recognition scores for words familiarised in happy but not in neutral emotion text passages (Singh et al., 2004). Words produced with an emotional $f_0$ assist infants in establishing representations and facilitate their word learning. While this does not automatically imply that they have decoded the emotional function carried on $f_0$, they are clearly sensitive to the $f_0$ differences between words produced with a neutral versus emotional $f_0$. Infants in their first year of life appear to have the capacity to separate linguistic and emotional functions of $f_0$, but no direct evidence of that has been reported.

Discussion on the interaction between intonational and emotional $f_0$ functions can be found in the area of infant-directed speech (IDS), a distinctive speech style that caretakers use to communicate with infants (Fernald, 1985, 1992). IDS is more exaggerated, with higher $f_0$ and wider $f_0$ ranges than adult-directed speech (ADS). Infants prefer IDS over ADS across the world's languages (ManyBabies Consortium, 2020). Some identify intonation as the key reason for this preference (Katz et al., 1996), whereas others attribute it to its attention-grabbing qualities (Burnham et al., 2002) and the positive emotion embedded in IDS (Singh et al., 2002). Infants appear to be sensitive to $f_0$ variations as early as 4 months of age, when they prefer $f_0$ but not amplitude or duration variations in IDS (Fernald and Kuhl, 1987). The fact that pragmatic functions encompassing both intonation and emotion, such as approval or prohibition, are more clearly expressed in IDS than in ADS, suggests that infants are capable of identifying those $f_0$ functions (Fernald, 1989; Moore et al., 1997). Indeed, as early as 5 months, infants are able to associate positive emotion in IDS with approval vocalisations, and negative emotion with prohibition vocalisations (Fernald, 1993). The functions of IDS appear to change over the first year of life, with ratings of mothers' IDS showing general decrease in comforting and soothing functions, and an increase in attentional and directive functions (Kitamura and Burnham, 2003). Infants' preferences for those functions appear to follow the same developmental trend (Kitamura and Lam, 2009). Despite infants' clear sensitivity to these $f_0$ patterns, another study suggests that children do not consider $f_0$ in speech as a reliable cue to indicate emotions until around 4–5 years of age (Quam and Swingley, 2012).

To our knowledge, no study has attempted to tease apart the three-way interaction between tone, intonation, and emotional functions in $f_0$. Trends may be observed in emotional $f_0$ from its immense variations, but not "rules" in the same sense as tone (e.g., "a tone language has a set of fixed pitch variations") or intonation (e.g., "a question usually has a rising pitch"). Thus, while there are broad indicators about the association between $f_0$ and emotion, this relationship, as well as its consistency across languages and cultures, is still under investigation. The interactions in between tone, intonation and emotion remain unclear, and research on IDS cannot efficiently disentangle its impact from intonational or emotional perspectives.

## Summary

The fluctuating $f_0$ signal contains overlapping information from different sources that infants need to decompose and recognise. We have focused on three distinct functions carried by $f_0$; tone, intonation, and emotion. It is not yet clear whether languages differ from each other in the way that emotion is expressed using $f_0$, but there are clear differences in the ways that languages use $f_0$ for tone and intonation. Infants do not know innately whether the information in $f_0$ refers to tone, intonation, or emotion. They must learn which aspects of the fluctuating $f_0$ signal correspond to different functions.

Studies on the developmental trajectories of infants' sensitivity to the tonal, intonational, and emotion aspects carried on $f_0$ have yielded mixed findings. Unstable and fluctuating developmental trajectories have been reported for tone, not only for infants learning a tone language but also for those learning a non-tone language in the first 2 years of life. Similarly, infants' intonation development does not appear to be linear before Year 2, and children do not use $f_0$ for intonation reliably until after Year 3. Although the contribution of $f_0$ on emotion is widely acknowledged, incongruent findings have been reported across the world's languages. Reliable use of $f_0$ as a cue to indicate emotion has only been found after Year 4 (Quam and Swingley, 2012).

Research on infant speech perception has only recently begun to focus on $f_0$ and there is certainly more work that needs to be done to establish clear developmental patterns. Nevertheless, it is clear that infants are sensitive to $f_0$ across domains, in tone (Liu and Kager, 2014), intonation (Frota et al., 2014) and emotion (Singh et al., 2004), and it appears that robust knowledge about tone is learned ahead of intonation and emotion. This observation is consistent with the idea that discrete categories for tone seem to be established earlier and more easily than they are for intonation (Tonkova-Yampol'skaya, 1969; Snow, 2006; Yeung et al., 2013). Indeed, it could be argued that the variability in the way that the three functions are represented in $f_0$ increases from tone, to intonation, then emotion. Such variability would make an infant's job of learning the $f_0$ patterns even more challenging, which may explain the developmental progression and fluctuation across domains.

Although traces of overlap in between these domains appear in literature, there is insufficient empirical data to disentangle the interactions between tone, intonation and emotion in the development of $f_0$ perception. To arrive at a clear explanation

of how infants learn to use $f_0$ cues in linguistic and paralinguistic functions, it is necessary to formulate a theoretical framework that incorporates $f_0$ functions across multiple domains.

## THEORETICAL CONSIDERATIONS

Investigating how infants solve the puzzle of decomposing $f_0$ into different functions is a rare opportunity to observe language development across different communicative domains. One interesting aspect of $f_0$, from a developmental perspective, is that an $f_0$ pattern that signals a tonal function in one language could be perceived as intonation in another. Proposing a perspective that can conceptually integrate across all three lines of inquiry – tone, intonation and emotion – may seem ambitious, but it is necessary to consider all of these aspects to understand how infants learn to decode $f_0$. Given the developmental patterns that have been observed for the three domains, a purely bottom-up statistical learning solution seems unlikely. Rather, infants may require multimodal experiences from their environment to develop functional speech communication skills. Our current understanding of how tone or intonation is coded in the visual modality, and how emotion is coded in the auditory speech signal is rudimentary. Nevertheless, addressing the multifunctionality of the speech signal using a global approach, conveying linguistic, paralinguistic, and affective information simultaneously, is critical for a comprehensive model of speech development. Any theory addressing $f_0$ perception and development will need to be able to explain how children acquire their native $f_0$ functions and account for the mixed findings observed in previous literature. On these bases, we argue for four critical aspects that must be properly addressed by any theories concerning $f_0$ perception and development.

- *Disambiguation*: how infants disentangle and recognise multiple overlapping $f_0$ patterns
- *Categorisation*: how infants learn that those patterns correspond to a given (native) linguistic or paralinguistic function
- *Accomodation*: how infants tackle $f_0$ functions that deviate from their native functional use
- *Interaction*: why recognition, learning and cue weighting of $f_0$ fluctuate along the development

Below, we consider how developmental theories of speech perception, cognition, and statistical learning may contribute to a broad theoretical approach to explaining the eventual successful acquisition of $f_0$ functions.

### Speech Perception

From a developmental perspective, *Perceptual Attunement* accounts (Werker and Hensch, 2015; Reh et al., 2020) propose that an infant's perception gradually shifts from universal into native or environmentally-attenuated perception patterns. Such changes occur across domains and modalities, fitting well in the aspect of *categorisation*. Such accounts associate well with, and arguably, lay the foundation of speech processing theories. For linguistic functions such as tone and intonation, infants typically exhibit initial biases or universal sensitivity, and quickly tune into the $f_0$ patterns of their native language (Burnham, 1986). Meanwhile, assimilations or perceptual difficulties surface since non-native or unfamiliar $f_0$ patterns are tuned out. Having said that, discrepancies from the attunement process have been reported for native and non-native $f_0$ patterns (Fikkert et al., 2020). Though overlapping $f_0$ patterns have been used as a possible explanation for these findings, theories of perceptual attunement will need to demonstrate *disambiguation*: how infants overcome overlaps in (e.g., $f_0$) functions along the developmental trajectory.

Further, models and theories of infants' acquisition of their L1 phonological system have been devised to explain how infants tune in to the phonetic features that signal phonological similarities and differences in the language of their environment (e.g., Best, 1994; Escudero, 2005; Kuhl et al., 2008; Polka and Bohn, 2011). The focus of these models has been on the acquisition of consonants and vowels (henceforth, *phones*). Here, we use the framework of the *Perceptual Assimilation Model* (PAM; Best, 1994; Best et al., 2009, Tyler et al., 2014) to consider how such models might account for the acquisition of $f_0$ functions.

A key empirical observation that led to the development of PAM was that English infants and adults had high discrimination accuracy for non-native Zulu click consonants despite never having encountered them before (Best et al., 1988). When asked to write down what they heard, all participants reported relying on non-speech characteristics of the consonants (e.g., water dripping, fingers snapping, or tongue popping). To account for this, PAM proposes that non-native phones may be perceived as speech (i.e., assimilated to the native phonological system) or as non-speech. When perceived as speech, a non-native phone may be assimilated as categorised (as a good, medium, or poor exemplar of a native phonological category) or uncategorised (not a clear exemplar of any single L1 category). Discrimination of non-native phonemes that are perceived as speech is crucially dependent on how it is assimilated to the native phonological system. Sometimes natively tuned perception will support discrimination (e.g., when each non-native phone is assimilated to a different L1 phonological category) and sometimes it will make it difficult to perceive any differences between them (e.g., when the non-native phones are perceived as equally good or poor exemplars of the same L1 category). Contrasting non-native phones that are perceived as non-speech (e.g., click consonants) are discriminated well by adults because they learned that the phonetic features of these categories are not used for linguistic purposes in their native language. Consistent with this account, native speakers of the click languages Zulu and Sesothu predominantly perceived non-native!Xóõ click consonants as speech (Best et al., 2003). Both click consonants in one of the !Xóõ contrasts were perceived as the same L1 click consonant category by both Zulu and Sesothu listeners. Importantly, English listeners perceived the same click consonants predominantly as non-speech and their discrimination of the contrast was more accurate than both groups of click language speakers. It appears that

the English speaking adults had learned, as infants, that the phonetic characteristics that correspond to click consonants were not part of the L1 phonological space.

According to PAM, infants transition from language-independent phonetic sensitivity to natively tuned perception by recognising higher order invariant information in articulatory patterns through processes of perceptual learning (Gibson and Pick, 2000). Phonetic variability is crucial for phonological development because infants need to learn not only those phonetic differences that signal a difference in meaning (the principle of phonological distinctiveness), but also those variable phonetic characteristics that define a category (the principle of phonological constancy; Best et al., 2009; Best, 2015). The region of phonetic space that is dedicated to speech is known as the phonological space. Click consonants would fall outside of the phonological space for English speakers but they would fall inside the phonological space for click language speakers. The development of phonological categories is beneficial for L1 perception because it supports accurate and rapid detection of the critical phonetic differences that signal a potential difference in word meaning. However, once infants have begun to tune into the L1 phonology, non-native speech is also perceived in terms of its similarities and differences to their developing L1 phonological categories. If they happen to perceive each phoneme in a non-native contrast as different L1 phonological category (e.g., one phoneme as /b/and the other as/d/, a PAM two-category assimilation) then their natively tuned perception will still support rapid and accurate discrimination. However, if both non-native phonemes are perceived as the same L1 category (e.g., the Hindi dental vs. retroflex plosive contrast for English native speakers, Werker and Logan, 1985; a PAM single-category assimilation), then discrimination is poor.

If fluctuating $f_0$ patterns were considered in a similar way as the varying articulatory-acoustic patterns that demarcate consonants and vowels, then it is conceivable that infants might use similar learning mechanisms to separate the linguistic, paralinguistic, or extralinguistic functions carried on $f_0$. For example, the $f_0$ patterns that are used in a tone language for lexical distinctions may be similar to those used for other functions in a non-tone language, such as intonation (for a discussion, see, Best, 2019). The developmental changes in infants' responses to $f_0$ fluctuation might then be explained by infants' learning and recognition of the various functions at different ages. For infants who experience phonological characteristics of a non-tone language, $f_0$ is irrelevant for lexical distinctions. This may explain why discrimination of tonal contrasts initially declines. The subsequent improvement would then be due to the development of sensitivity to other types of $f_0$ information. Thus, from the perspective of the Perceptual Assimilation Model, *disambiguation* and *categorisation* occur through processes of perceptual learning. *Accommodation* may be observed if infants perceive a non-native $f_0$ pattern as consistent with a different type of function in their L1, and *interaction* may be explained by the different timescales for perceptual development of linguistic, paralinguistic, and extralinguistic information.

## Cognition

Another potential joinder of the three areas of $f_0$ functions resides in cognitive competition. Theories such as the *Functional Load Hypothesis* (FLH, Berinstein, 1979) postulate that our prosodic space of a given language is finite, and therefore, assume competition in phonological processing. Under FLH, it would be more cognitively demanding to process $f_0$ contours that simultaneously carry more than one type of function.

The FLH predictions provide indirect explanations for *disambiguation,* as presumably, competition across diverse $f_0$ functions may facilitate their recognition, disentanglement and establishment of $f_0$ categories. These predictions also offer viable ways of empirically examining FLH as a hypothesis. Having said that, existing findings are mixed (van Heuven, 2018). FLH is supported by studies investigating parameters competing within the prosodic domain. Supported by phonological and acoustic analyses, Remijsen (2002) has shown that it is highly unlikely for a tone language to feature lexical stress because that would create competition (and thus ambiguity) between the pragmatic and the lexical functions of $f_0$. Using phonological and acoustic analyses, Remijsen (2002) concluded that it is implausible for lexical tone and lexically contrastive stress accent to co-exist in the word-prosodic system of a language. Nevertheless, challenges appear to lie in *interaction*: FLH would need to explain how parameters from different domains within phonology (e.g., prosodic vs. segmental domains) and beyond (e.g., linguistic vs. paralinguistic domains) compete against one another. In other words, it is unclear whether and to what extent information across domains and modalities fights for cognitive resources during processing. FLH concentrates on the linguistic domain and the emotional aspect has not been directly considered (although it was alluded to in Chen, 2005). Nevertheless, the FLH postulation seems to imply that languages encoding $f_0$ in both tone and intonation would have less functional space left to encode $f_0$ in emotions. Note that caregivers may assist, consciously or unconsciously, in the reduction of functional loads in the course of infants' learning. For instance, they may package messages in IDS to reduce processing challenges for certain $f_0$ functions.

The FLH faces challenges incorporating cross-domain or cross-modal facilitation effects. That is, information perceived in one domain (e.g., vision) may support perception and learning of information in another domain (e.g., speech). These are often referred to as bootstrapping or anchoring effects. For instance, the prosodic bootstrapping hypothesis suggests that infants may use prosodic information to discover utterance and word boundaries (Seidl and Johnson, 2006; Johnson et al., 2014), and knowledge of word semantics may further cue syntactic categories (Höhle, 2009). Along the same lines, various sources of information from the ambient environment provide anchors to facilitate children's $f_0$ *disambiguation* and *categorisation* along the developmental trajectory. The command of one $f_0$ function may facilitate another even when they are simultaneously presented. FLH, or any cognitive model, will need to clearly explain the degree of interaction between competition and facilitation in co-occurring functions.

What has not been discussed, but links closely with the FLH mechanisms, is how infants cope with cognitive demands and how increased neurocognitive ability affects children's perception and learning. It takes children years to master linguistic and pragmatic functions. Taking *Theory of Mind* (ToM) as an example, ToM refers to the understanding of distinctions between individuals' mental states, mental constructs, physical entities and their overt actions (Gopnik and Wellman, 1992; Wellman, 1992). ToM is crucial for children's socio-emotional development. What needs to be explored is how children's gross and specific (e.g., socio-emotions) cognitive development attributes to the learning of emotional $f_0$.

## Statistical Learning

Statistical learning refers to the ability to acquire information solely based on relevant statistical distributions in the ambient environment, and *Statistical Learning* accounts argue that infants utilise their innate statistical (Saffran and Kirkham, 2018) and relational (Ferry et al., 2015) learning ability to acquire new information. For instance, 8-month-olds are able to segment words from fluent speech based on one and only one cue: the statistical relationships between neighbouring syllables (Saffran et al., 1996; but see Johnson and Tyler, 2010).

While statistical learning accounts have been used to describe acquisition of a single $f_0$ function (e.g., lexical tone, Liu and Kager, 2017b), its explanatory power faces evident challenges in *disambiguation* and *categorisation*. A purely bottom-up learning of a statistical distribution does not appear sufficient to explain *disambiguation* if $f_0$ is the only statistical distribution available. By comparison, vowels may be disambiguated on the basis of multiple information sources (e.g., the first, second, and third formants, and duration). Even though $f_0$ serves as the primary acoustic correlate of emotional tones (Scherer, 2003), its usage differs between tone and non-tone language speakers, with greater $f_0$ variations in the productions of the latter group. It seems likely that statistical learning of $f_0$ patterns would require correlated statistical distributions from other information sources. This may include phonation type (e.g., creaky voice) or tone-vowel interactions (Shaw and Tyler, 2020) for tone, and voice quality for both intonation (Ladd et al., 1985) and emotion (Yanushevskaya et al., 2018). Cue-weighting, or differences in listeners' weighting of acoustic cues (e.g., between $f_0$ and secondary cues such as amplitude and duration, Ross et al., 1986), likely further modulates statistical learning.

With respect to *categorisation*, statistical learning ability does not appear to be constant across ages. Its efficacy changes dynamically over a child's development. However, the direction of such change, or the statistical learning efficacy across ages, is currently a matter of debate. On the one hand, a meta-analysis has reported increased effect sizes with age in the first year of life (Cristia, 2018), suggesting that older infants are increasingly sensitive to this learning mechanism. On the other hand, behavioural (Yoshida et al., 2010) and neural (Wanrooij et al., 2014) evidence has shown that this learning mechanism may be maturationally delimited, along the perceptual attunement trajectory during which phonetic perception is refined (Liu and Kager, 2017b; Reh et al., 2021). The latter evidence suggests that the learning of sound frequency distributions become increasingly resistant as children grow. Discrepancies in literature have been explained by the different perceptual attunement time windows of speech sounds differing in phonetic representations, space and perceptual/acoustic salience (Werker and Hensch, 2015; Reh et al., 2021). Hence, statistical learning of speech sounds may be at its peak of efficacy during perceptual attunement, when infants' perception exhibits enhanced sensitivity to input from the environment.

Although the learning mechanism is considered domain- (and even species-) general, individual studies and models typically investigate statistical learning in a domain-specific fashion. Despite the challenge in *disambiguation* and the debate in *categorisation*, in order to achieve learning of diverse $f_0$ functions, models of statistical learning would require additional focus on the *interaction* mechanisms, with modelling of certain (e.g., $f_0$) statistical distributions across domains.

## Summary

Similar to the lack of empirical research in studying the interaction of distinct linguistic and paralinguistic functions carried on $f_0$, none of the existing models and hypotheses seems sufficient in addressing how different $f_0$ functions disassociate in sensory and cognitive processes, or the extent to which they are processed simultaneously or separately. A theoretical account is required for how infants manage to decompose these overlapping $f_0$ functions while taking into consideration the differences between these functions across languages/cultures, as well as information integration across modalities.

As summarised in the beginning of this section, to achieve successful learning, infants must rely on fundamental aspects (disambiguation, categorisation, accommodation and interaction). These aspects point out directions where the exploration of diverse $f_0$ functions may converge. These directions are crucial for us to understand how infants resolve puzzles identified in the literature:

- *Neuro-cognitive Development*, which reflects age-related developmental and maturational changes
- *Environmental Information*, where learning of language and social-emotions from the ambient resources occur
- *Competition and Facilitation*, within and across perceptual and/or cognitive spaces and modalities (e.g., auditory, visual) where information gathers and integrates

Infants eventually sort out their native linguistic and socio-emotional functions carried on $f_0$. Thus, developmental and environmental aspects such as age and experience will need to be considered when exploring $f_0$ functions, in line with the first two directions. With respect to category learning, future research should focus on the establishment of $f_0$ categories for tone, intonation, and emotion. Further that, the degree of flexibility and assimilation when facing a novel/non-native category will need to be explored. Regarding bootstrapping, a theoretical basis will require that infants effectively integrate environmental sources of information and existing knowledge

to recognise and disambiguate $f_0$ functions.[1] A multimodal view into the issue is also consistent with an ecological approach to perceptual learning and development (e.g., Gibson and Pick, 2000).

To summarise across the four directions, future research should concentrate on how infants decompose and acquire linguistic and paralinguistic functions carried on $f_0$; to what extent reinforcement or interference may occur with infants' perception and learning of $f_0$ functions; and how infants employ environmental resources to disambiguate these functions.

## Our Hypothesis

Considering the gap in discussion of $f_0$ functions across linguistic and socio-emotional domains, the four aspects concerning $f_0$ perception and development, and the four directions essential to achieve its functional learning, we propose a *Learnability Hypothesis* that infants require multimodal environmental experiences to decompose and acquire overlapping linguistic, paralinguistic, and extralinguistic $f_0$ functions. Its predictions are as follows: When faced with $f_0$ contours carrying multiple functions, perception and learning of a certain function should be enhanced if other functions are not ambiguous, and should be affected if other functions have not been properly learned or cannot be properly identified. Moreover, infants use acquired, environmental and multi-modal cues to anchor and facilitate learning whenever possible.

A representative and measurable case of the learnability hypothesis can be viewed as the "synergy scenario." For example, infants can use visual cues to disambiguate and decompose different auditory $f_0$ functions. Congruent audiovisual cues of the same function will lead to corresponding enhancements as well as reduced sensitivity to others. In contrast, incongruent cues may capture infants' attention, as is the case for deviants against standards in an oddball paradigm in electroencephalogram, or regained attention to new information in a behavioural habituation paradigm. These predictions provide us with viable ways of testing the hypothesis.

One way to examine this scenario would be to use an experimental paradigm that reflects the real world lives and interactions that infants experience, such as using stimuli that mirror real communications that occur in infant-caregiver interactions. Following an associative learning paradigm (Hay et al., 2015; Liu and Kager, 2018), infants' ability to associate novel objects with an instructor's $f_0$s that represent tones could be measured with or without the instructor's visual intonational and emotional information. Here, the $f_0$s could be ambiguous, not only reflecting tonal but also intonational or emotional $f_0$ that are relevant in infants' native environment. In this case, when the presented visual information matched intonational or emotional $f_0$, infants should show a reduction in associative learning.

[1]Recent evidence suggests that there may be information about $f_0$ in the face and in head movements that can be used to discriminate lexical tone contrasts (Burnham et al., 2022), but it is not clear from these findings whether such auditory visual speech information would be useful for disambiguating different $f_0$ functions. Here we consider the role of non-speech environmental information on the acquisition of $f_0$ functions.

# CONCLUSION

A diverse array of linguistic and paralinguistic functions are carried simultaneously on $f_0$. Patterns of $f_0$ variability differ across languages, such that an $f_0$ pattern that serves a particular function in one language may serve a different function in another. Adults use native $f_0$ functions effortlessly, but how infants acquire them remains a mystery. Infants' unstable learning trajectories raise important questions. For instance, when they no longer treat $f_0$ differences as potential signals to a change in a certain function, is it due to an insensitivity to $f_0$ features or due to those features being used for a different communicative purpose? Do infants adopt top-down or bottom-up processing when disambiguating different functions carried on the same $f_0$? These questions surface from the mixed findings in the literature, across tone, intonation, and emotional domains.

It is important to seek answers to these questions and solutions to the discrepancies observed in the literature. The body of literature needs to be expanded to include infants from a broader range of language environments so that we can understand the course of acquisition. Obtaining the answers through a theoretical and empirical approach, such as the research ideas spawned by our *Learnability Hypothesis*, will improve and integrate theories across research fields, especially when existing models do not appear sufficiently inclusive to address the learning process.

The early years of life lay solid foundations for child learning, assisting our young learners to navigate through the complexities of our modern world. The understanding of how children command the multiple $f_0$ functions using an ecological approach will function as a benchmark guiding pitch learning in the natural environment; help with the identification of speech or cognitive impairments; better support typical child development; and contribute to multilingual/vulnerable language learning, second/foreign language learning, as well as learning across the lifespan.

# AUTHOR CONTRIBUTIONS

LL and MT drafted and revised the manuscript. AG and PL revised the manuscript. All authors contributed to the article and approved the submitted version.

# FUNDING

LL's writing was partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 798658 hosted by Center for Multilingualism across the Lifespan at the University of Oslo, financed by Research Council of Norway through its Centers of Excellence funding scheme grant agreement No. 223265. The open access publication fee was supported by Western Sydney University.

## REFERENCES


Anolli, L., Wang, L., Mantovani, F., and De Toni, A. (2008). The voice of emotion in Chinese and Italian young adults. *J. Cross-Cult. Psychol.* 39, 565–598. doi: 10.1177/0022022108321178

Arvaniti, A., and Fletcher, J. (2020). "The autosegmental-metrical theory of intonational phonology," in *The Oxford Handbook of Language Prosody*. eds. C. Gussenhoven and A. Chen (Oxford: Oxford University Press), 78–95.

Banse, R., and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol.* 70, 614–636. doi: 10.1037/0022-3514.70.3.614

Bänziger, T., Hosoya, G., and Scherer, K. R. (2015). Path models of vocal emotion communication. *PLoS One* 10:e0136675. doi: 10.1371/journal.pone.0136675

Belyk, M., and Brown, S. (2014). Perception of affective and linguistic prosody: An ALE meta-analysis of neuroimaging studies. *Soc. Cogn. Affect. Neurosci.* 9, 1395–1403. doi: 10.1093/scan/nst124

Berinstein, A. E. (1979). A cross-linguistic Study on the Perception and Production of Stress Unpublished master's dissertation. University of California Los Angeles.

Best, C. T. (1994). "The emergence of native-language phonological influences in infants: A perceptual assimilation model," in *The Development of speech Perception: The transition from speech Sounds to Spoken Words*. eds. J. C. Goodman and H. C. Nusbaum (United States: MIT Press), 167–244.

Best, C. T. (2015). "Devil or angel in the details? Perceiving phonetic variation as information about phonological structure," in *Phonetics-phonology interface: Representations and methodologies*. eds. J. Romero and M. Riera (Amsterdam: John Benjamins), 3–31.

Best, C. T. (2019). The diversity of tone languages and the roles of pitch variation in non-tone languages: considerations for tone perception research. *Front. Psychol.* 10:364. doi: 10.3389/fpsyg.2019.00364

Best, C. T., McRoberts, G. W., and Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *J. Exp. Psychol. Hum. Percept. Perform.* 14, 345–360. doi: 10.1037/0096-1523.14.3.345

Best, C. T., Traill, A., Carter, A., Harrison, K. D., and Faber, A. (2003). "!Xóõ Click Perception by English, Isizulu, and Sesotho Listeners." in *Proceedings of the 15th International Congress of Phonetic Sciences*. eds. M. J. Solé, D. Recasens and J. Romero; August 3–9, 2003; Causal Productions; 853–856.

Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., and Quann, C. A. (2009). Development of phonological constancy: Toddlers' perception of native-and Jamaican-accented words. *Psychol. Sci.* 20, 539–542. doi: 10.1111/j.1467-9280.2009.02327.x

Braun, B., and Johnson, E. K. (2011). Question or tone 2? How language experience and linguistic function guide pitch processing. *J. Phon.* 39, 585–594. doi: 10.1016/j.wocn.2011.06.002

Bryant, G. A. (2021). The evolution of human vocal emotion. *Emot. Rev.* 13, 25–33. doi: 10.1177/1754073920930791

Burnham, D. K. (1986). Developmental loss of speech perception: exposure to and experience with a first language1. *Appl. Psycholinguist.* 7, 207–239. doi: 10.1017/S0142716400007542

Burnham, D., Kitamura, C., and Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science* 296:1435. doi: 10.1126/science.1069587

Burnham, D. K., and Singh, L. (2018). Coupling tonetics and perceptual attunement: The psychophysics of lexical tone contrast salience. *J. Acoust. Soc. Am.* 144:1716. doi: 10.1121/1.5067611

Burnham, D., Vatikiotis-Bateson, E., Barbosa, A. V., Menezes, J. V., Yehia, H. C., Morris, R. H., et al. (2022). Seeing lexical tone: head and face motion in production and perception of Cantonese lexical tones. *Speech Comm.* 141, 40–55. doi: 10.1016/j.specom.2022.03.011

Cahn, J. E. (1990). The generation of affect in synthesized speech. *J. Am. Voice I/O Soc.* 8:1

Caron, R. F., Caron, A. J., and Myers, R. S. (1985). Do infants see emotional expressions in static faces? *Child Dev.* 56, 1552–1560. doi: 10.2307/1130474

Chao, Y. R. (1947). *Cantonese Primer*. United States: Harvard University Press.

Chen, A. (2005). *Universal and Language-specific Perception of Paralinguistic Intonational Meaning*. Utrecht: LOT.

Chen, A., and Kager, R. (2016). Discrimination of lexical tones in the first year of life. *Infant Child Dev.* 25, 426–439. doi: 10.1002/icd.1944

Chen, L. M., and Kent, R. D. (2009). Development of prosodic patterns in mandarin-learning infants. *J. Child Lang.* 36, 73–84. doi: 10.1017/S0305000908008878

Chen, A., Liu, L., and Kager, R. (2015). Cross-linguistic perception of mandarin tone sandhi. *Lang. Sci.* 48, 62–69. doi: 10.1016/j.langsci.2014.12.002

Chen, A., Stevens, C. J., and Kager, R. (2017). Pitch perception in the first year of life, a comparison of lexical tones and musical pitch. *Front. Psychol.* 8:297. doi: 10.3389/fpsyg.2017.00297

Chien, P. J., Friederici, A. D., Hartwigsen, G., and Sammler, D. (2020). Neural correlates of intonation and lexical tone in tonal and non-tonal language speakers. *Hum. Brain Mapp.* 41, 1842–1858. doi: 10.1002/hbm.24916

Chong, S. C. F., Werker, J., Russell, J. A., and Carroll, J. M. (2003). Three facial expressions mothers direct to their infants. *Infant Child Dev.* 12, 211–232. doi: 10.1002/icd.286

Cristia, A. (2018). Can infants learn phonology in the lab? A meta-analytic answer. *Cognition* 170, 312–327. doi: 10.1016/j.cognition.2017.09.016

Crystal, D., and Quirk, R. (1964). *Systems of Prosodic and Paralinguistic Features in English*. Berlin: De Gruyter Mouton.

Dachkovsky, S., and Sandler, W. (2009). Visual intonation in the prosody of a sign language. *Lang. Speech* 52, 287–314. doi: 10.1177/0023830909103175

Dijkstra, C., Krahmer, E., and Swerts, M. (2006). "Manipulating uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence." in *Proceedings of the Speech Prosody Conference, Dresden*. May 2–5, 2006.

Doan, S. N. (2010). The role of emotion in word learning. *Early Child Dev. Care* 180, 1065–1078. doi: 10.1080/03004430902726479

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: towards a new generation of databases. *Speech Comm.* 40, 33–60. doi: 10.1016/S0167-6393(02)00070-5

Escudero, P. (2005). *Linguistic Perception and second Language Acquisition: Explaining the Attainment of Optimal Phonological Categorization*. Netherlands: Netherlands Graduate School of Linguistics.

Esteve-Gibert, N., Prieto, P., and Liszkowski, U. (2017). Twelve-month-olds understand social intentions based on prosody and gesture shape. *Infancy* 22, 108–129. doi: 10.1111/infa.12146

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behav. Dev.* 8, 181–195. doi: 10.1016/S0163-6383(85)80005-9

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: is the melody the message? *Child Dev.* 60, 1497–1510. doi: 10.2307/1130938

Fernald, A. (1992). "Human maternal vocalizations to infants as biologically relevant signals," in *The Adapted mind: Evolutionary Psychology and the Generation of Culture*. eds. J. Barkow, L. Cosmides and J. Tooby (Oxford, England: Oxford University Press), 391–428.

Fernald, A. (1993). Approval and disapproval: infant responsiveness to vocal affect in familiar and unfamiliar languages. *Child Dev.* 64, 657–674. doi: 10.2307/1131209

Fernald, A., and Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behav. Dev.* 10, 279–293. doi: 10.1016/0163-6383(87)90017-8

Fernald, A., and Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Dev. Psychol.* 20, 104–113. doi: 10.1037/0012-1649.20.1.104

Ferry, A. L., Hespos, S. J., and Gentner, D. (2015). Prelinguistic relational concepts: investigating analogical processing in infants. *Child Dev.* 86, 1386–1405. doi: 10.1111/cdev.12381

Fikkert, P., Liu, L., and Ota, M. (2020). "The acquisition of word prosody," in *The Oxford Handbook of Language Prosody* (London: Oxford University Press), 541–552.

Francis, A. L., Ciocca, V., Ma, L., and Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *J. Phon.* 36, 268–294. doi: 10.1016/j.wocn.2007.06.005

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychol. Bull.* 97, 412–429. doi: 10.1037/0033-2909.97.3.412

Fromkin, V. A. (Ed.). (2014). *Tone: A linguistic Survey*. United States: Academic Press.

Frota, S., and Butler, J. (2018). "Early development of intonation," in *The Development of Prosody in first Language Acquisition*. eds. P. Prieto and N. Esteve-Gibert (Netherland: John Benjamins), 145–164.

Frota, S., Butler, J., and Vigário, M. (2014). Infants' perception of intonation: is it a statement or a question? *Infancy* 19, 194–213. doi: 10.1111/infa.12037

Frühholz, S., Ceravolo, L., and Grandjean, D. (2012). Specific brain networks during explicit and implicit decoding of emotional prosody. *Cereb. Cortex* 22, 1107–1117. doi: 10.1093/cercor/bhr184

Gandour, J. (1983). Tone perception in far eastern languages. *J. Phon.* 11, 149–175. doi: 10.1016/S0095-4470(19)30813-7

Geangu, E., Ichikawa, H., Lao, J., Kanazawa, S., Yamaguchi, M. K., Caldara, R., et al. (2016). Culture shapes 7-month-olds' perceptual strategies in discriminating facial expressions of emotion. *Curr. Biol.* 26, R663–R664. doi: 10.1016/j.cub.2016.05.072

Gendron, M., Crivelli, C., and Barrett, L. F. (2018). Universality reconsidered: diversity in making meaning of facial expressions. *Curr. Dir. Psychol. Sci.* 27, 211–219. doi: 10.1177/0963721417746794

Gendron, M., Roberson, D., and Barrett, L. F. (2015). Cultural variation in emotion perception is real: A response to Sauter, Eisner, Ekman, and Scott (2015). *Psychol. Sci.* 26, 357–359. doi: 10.1177/0956797614566659

Gendron, M., Roberson, D., van der Vyver, J. M., and Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion* 14, 251–262. doi: 10.1037/a0036052

Gibson, E. J., and Pick, A. D. (2000). *An Ecological Approach to Perceptual Learning and Development*. England: Oxford University Press.

Gobl, C., Bennett, E., and Chasaide, A. N. (2002). Expressive synthesis: how crucial is voice quality?. In *Proceedings of 2002 IEEE Workshop on Speech Synthesis*. September 11–13, 2002; IEEE; 91–94.

Gobl, C., and Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Comm.* 40, 189–212. doi: 10.1016/S0167-6393(02)00082-1

Goldbeck, T., Tolkmitt, F., and Scherer, K. R. (1988). "Experimental studies on vocal affect communication," in *Facets of Emotion: Recent Research*. ed. K. R. Scherer (Mahwah: Lawrence Erlbaum), 119–137.

Gopnik, A., and Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind Lang.* 7, 145–171. doi: 10.1111/j.1468-0017.1992.tb00202.x

Götz, A., Yeung, H. H., Krasotkina, A., Schwarzer, G., and Höhle, B. (2018). Perceptual reorganization of lexical tones: effects of age and experimental procedure. *Front. Psychol.* 9:477. doi: 10.3389/fpsyg.2018.00477

Goudbeek, M., and Scherer, K. (2010). Beyond arousal: valence and potency/control cues in the vocal expression of emotion. *J. Acoust. Soc. Am.* 128, 1322–1336. doi: 10.1121/1.3466853

Graf Estes, K., and Hay, J. F. (2015). Flexibility in bilingual infants' word learning. *Child Dev.* 86, 1371–1385. doi: 10.1111/cdev.12392

Grice, M., Baumann, S., and Benzmüller, R. (2006). *Autosegmental-Metrical Phonology*. Prosodic typology: The phonology of intonation and phrasing, 55–83.

Grossmann, T., Striano, T., and Friederici, A. D. (2005). Infants' electric brain responses to emotional prosody. *Neuroreport* 16, 1825–1828. doi: 10.1097/01.wnr.0000185964.34336.b1

Gussenhoven, C. (2002). "Intonation and interpretation: phonetics and phonology." in *Proceedings of the 1st International Conference on Speech Prosody*. April 11–13, 2002; 47–57.

Gussenhoven, C. (2004). *The Phonology of tone and Intonation*. England: Cambridge University Press.

Gussenhoven, C., and Chen, A. (2000). Universal and language-specific effects in the perception of question intonation. In *6th International Conference on Spoken Language Processing (ICSLP 2000)* 91–94.

Haviland, J. M., and Lelwica, M. (1987). The induced affect response: 10-week-old infants' responses to three emotion expressions. *Dev. Psychol.* 23, 97–104. doi: 10.1037/0012-1649.23.1.97

Hay, J. F., Cannistraci, R. A., and Zhao, Q. (2019). Mapping non-native pitch contours to meaning: perceptual and experiential factors. *J. Mem. Lang.* 105, 131–140. doi: 10.1016/j.jml.2018.12.004

Hay, J. F., Graf Estes, K., Wang, T., and Saffran, J. R. (2015). From flexibility to constraint: The contrastive use of lexical tone in early word learning. *Child Dev.* 86, 10–22. doi: 10.1111/cdev.12269

He, C., Hotson, L., and Trainor, L. J. (2007). Mismatch responses to pitch changes in early infancy. *J. Cogn. Neurosci.* 19, 878–892. doi: 10.1162/jocn.2007.19.5.878

Hoemann, K., Xu, F., and Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Dev. Psychol.* 55, 1830–1849. doi: 10.1037/dev0000686

Höhle, B. (2009). Bootstrapping mechanisms in first language acquisition. *Linguistics* 47, 359–382. doi: 10.1515/LING.2009.013

Hua, Z. (2002). *Phonological Development in specific Contexts: Studies of Chinese-Speaking children. Vol. 3* United Kingdom: Multilingual Matters.

Hua, Z., and Dodd, B. (2000). The phonological acquisition of Putonghua (modern standard Chinese). *J. Child Lang.* 27, 3–42. doi: 10.1017/S030500099900402X

Hyman, L. M. (2011). "Tone: is it different?" in *The Handbook of Phonological Theory. Vol. 75*. eds. J. A. Goldsmith, J. Riggle and A. C. L. Yu (United States: John Wiley and Sons), 50–80.

Hyman, L. M. (2016). Lexical vs. grammatical tone: sorting out the differences. In *proceedings of the 5th international symposium on tonal aspects of languages (TAL 2016)*. May 24–27, 2016; 6–11.

Ilie, G., and Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music. Percept.* 23, 319–330. doi: 10.1525/mp.2006.23.4.319

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *J. Acoust. Soc. Am.* 88, 642–654. doi: 10.1121/1.399767

Johnson, E. K., Seidl, A., and Tyler, M. D. (2014). The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PLoS One* 9:e83546. doi: 10.1371/journal.pone.0083546

Johnson, E. K., and Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Dev. Sci.* 13, 339–345. doi: 10.1111/j.1467-7687.2009.00886.x

Jongman, A., Wang, Y., Moore, C., and Sereno, J. (2006). "Perception and production of mandarin tone," in *Handbook of East Asian Psycholinguistics. Vol. 1*. eds. P. Li, L. H. Tan, E. Bates and O. J. L. Tzeng (England: Cambridge University Press), 209–217.

Katz, G. S., Cohn, J. F., and Moore, C. A. (1996). A combination of vocal f0 dynamic and summary features discriminates between three pragmatic categories of infant-directed speech. *Child Dev.* 67, 205–217. doi: 10.1111/j.1467-8624.1996.tb01729.x

Keitel, A., Prinz, W., Friederici, A. D., Von Hofsten, C., and Daum, M. M. (2013). Perception of conversations: The importance of semantics and intonation in children's development. *J. Exp. Child Psychol.* 116, 264–277. doi: 10.1016/j.jecp.2013.06.005

Kent, R. D., and Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *J. Acoust. Soc. Am.* 72, 353–365. doi: 10.1121/1.388089

Kitamura, C., and Burnham, D. (2003). Pitch and communicative intent in mother's speech: adjustments for age and sex in the first year. *Infancy* 4, 85–110. doi: 10.1207/S15327078IN0401_5

Kitamura, C., and Lam, C. (2009). Age-specific preferences for infant-directed affective intent. *Infancy* 14, 77–100. doi: 10.1080/15250000802569777

Koeda, M., Belin, P., Hama, T., Masuda, T., Matsuura, M., and Okubo, Y. (2013). Cross-cultural differences in the processing of non-verbal affective vocalizations by Japanese and Canadian listeners. *Front. Psychol.* 4:105. doi: 10.3389/fpsyg.2013.00105

Kotz, S. A., Meyer, M., and Paulmann, S. (2006). Lateralization of emotional prosody in the brain: An overview and synopsis on the impact of study design. *Prog. Brain Res.* 156, 285–294. doi: 10.1016/S0079-6123(06)56015-7

Kotz, S. A., and Paulmann, S. (2007). When emotional prosody and semantics dance cheek to cheek: ERP evidence. *Brain Res.* 1151, 107–118. doi: 10.1016/j.brainres.2007.03.015

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., and Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Phil. Trans. R. Soc. B: Biol. Sci.* 363, 979–1000. doi: 10.1098/rstb.2007.2154

Ladd, D. R. (2008). *Intonational Phonology, 2nd Edn*. Cambridge: Cambridge University Press.

Ladd, D. R., Silverman, K. E., Tolkmitt, F., Bergmann, G., and Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *J. Acoust. Soc. Am.* 78, 435–444. doi: 10.1121/1.392466

Laukka, P., Juslin, P., and Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognit. Emot.* 19, 633–653. doi: 10.1080/02699930441000445

Lee, C. C., Jhang, Y., Chen, L. M., Relyea, G., and Oller, D. K. (2017). Subtlety of ambient-language effects in babbling: a study of English-and Chinese-learning infants at 8, 10, and 12 months. *Lang. Learn. Dev.* 13, 100–126. doi: 10.1080/15475441.2016.1180983

Lee, C. Y., and Lee, Y. F. (2010). Perception of musical pitch and lexical tones by mandarin-speaking musicians. *J. Acoust. Soc. Am.* 127, 481–490. doi: 10.1121/1.3266683

Li, A. (2015). *Encoding and Decoding of Emotional speech: A cross-Cultural and Multimodal Study between Chinese and Japanese*. United States: Springer.

Li, C. N., and Thompson, S. A. (1977). The acquisition of tone in mandarin-speaking children. *J. Child Lang.* 4, 185–199. doi: 10.1017/S0305000900001598

Liang, J., and Heuven, V. J. (2007). Chinese tone and intonation perceived by L1 and L2 listeners. In C. Gussenhoven and T. Riad (Eds.), *Tones and Tunes, Vol. 2: Experimental Studies in Word and Sentence Prosody*. Berlin: De Gruyter Mouton. (27–62)

Liebenthal, E., Silbersweig, D. A., and Stern, E. (2016). The language, tone and prosody of emotions: neural substrates and dynamics of spoken-word emotion perception. *Front. Neurosci.* 10:506. doi: 10.3389/fnins.2016.00506

Liu, L., du Toit, M., and Weidemann, G. (2021). Infants are sensitive to cultural differences in emotions at 11 months. *PLoS One* 16:e0257655. doi: 10.1371/journal.pone.0257655

Liu, L., and Kager, R. (2014). Perception of tones by infants learning a non-tone language. *Cognition* 133, 385–394. doi: 10.1016/j.cognition.2014.06.004

Liu, L., and Kager, R. (2015). Bilingual exposure influences infant VOT perception. *Infant Behavior and Development* 38, 27–36.

Liu, L., and Kager, R. (2017a). Perception of tones by bilingual infants learning non-tone languages. *Biling. Lang. Congn.* 20, 561–575. doi: 10.1017/S1366728916000183

Liu, L., and Kager, R. (2017b). Statistical learning of speech sounds is most robust during the period of perceptual attunement. *J. Exp. Child Psychol.* 164, 192–208. doi: 10.1016/j.jecp.2017.05.013

Liu, L., and Kager, R. (2018). Monolingual and bilingual infants' ability to use non-native tone for word learning deteriorates by the second year after birth. *Front. Psychol.* 9:117. doi: 10.3389/fpsyg.2018.00117

Liu, L., Lai, R., Singh, L., Kalashnikova, M., Wong, P. C. M., Kasisopa, B., et al. (2022). The tone atlas of perceptual discriminability and perceptual distance: Four tone languages and five language groups. *Brain Lang.* 229:105106. doi: 10.1016/j.bandl.2022.105106

Liu, P., and Pell, M. D. (2012). Recognizing vocal emotions in mandarin Chinese: a validated database of Chinese vocal emotional stimuli. *Behav. Res. Methods* 44, 1042–1051. doi: 10.3758/s13428-012-0203-3

Liu, L., Varghese, P., and Weidemann, G. (2019). "A bilingual advantage in infant pitch processing." in *Proceedings of the 19th International Congress of Phonetic Sciences*. August 5–9, 2019; International Phonetic Association; 1397–1401.

Liu, L., and Xu, R. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* 62, 70–87.

Maddieson, I. (2013). "Tone" in *The World Atlas of Language Structures Online*. eds. M. S. Dryer, S. Matthew and M. Haspelmath (Germany: Max Planck Institute for Evolutionary Anthropology).

Mampe, B., Friederici, A. D., Christophe, A., and Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Curr. Biol.* 19, 1994–1997. doi: 10.1016/j.cub.2009.09.064

Manfredi, C., Viellevoye, R., Orlandi, S., Torres-García, A., Pieraccini, G., and Reyes-García, C. A. (2019). Automated analysis of newborn cry: relationships between melodic shapes and native language. *Biomed. Sig. Proces. Cont.* 53:101561. doi: 10.1016/j.bspc.2019.101561

ManyBabies Consortium (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Adv. Methods Pract. Psychol. Sci.* 3, 24–52. doi: 10.1177/2515245919900809

Massaro, D. W., Cohen, M. M., and Tseng, C. Y. (1985). The evaluation and integration of pitch height and pitch contour in lexical tone perception in mandarin Chinese. *J. Chinese Ling.* 267–289.

Mattock, K., and Burnham, D. (2006). Chinese and English infants' tone perception: evidence for perceptual reorganization. *Infancy* 10, 241–265. doi: 10.1207/s15327078in1003_3

Mattock, K., Molnar, M., Polka, L., and Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition* 106, 1367–1381. doi: 10.1016/j.cognition.2007.07.002

Moore, D. S., Spence, M. J., and Katz, G. S. (1997). Six-month-olds' categorization of natural infant-directed utterances. *Dev. Psychol.* 33, 980–989. doi: 10.1037/0012-1649.33.6.980

Mozziconacci, S. J. L. (1998). *Speech Variability and Emotion: Production and Perception*. Netherlands: Technical University Eindhoven.

Murray, I. R., and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *J. Acoust. Soc. Am.* 93, 1097–1108. doi: 10.1121/1.405558

Nazzi, T., Floccia, C., and Bertoncini, J. (1998). Discrimination of pitch contours by neonates. *Infant Behav. Dev.* 21, 779–784. doi: 10.1016/S0163-6383(98)90044-3

Nelson, C. A., and De Haan, M. (1996). Neural correlates of infants' visual responsiveness to facial expressions of emotion. *Dev. Psychobiol.* 29, 577–595. doi: 10.1002/(SICI)1098-2302(199611)29:7<577::AID-DEV3>3.0.CO;2-R

Ortony, A., Clore, G. L., and Collins, A. (1990). *The Cognitive Structure of Emotions*. England: Cambridge University Press.

Papoušek, M. (1989). Determinants of responsiveness to infant vocal expression of emotional state. *Infant Behav. Dev.* 12, 507–524. doi: 10.1016/0163-6383(89)90030-1

Pell, M. D., and Kotz, S. A. (2011). On the time course of vocal emotion recognition. *PLoS One* 6:e27256. doi: 10.1371/journal.pone.0027256

Pell, M. D., Monetta, L., Paulmann, S., and Kotz, S. A. (2009). Recognizing emotions in a foreign language. *J. Nonverbal Behav.* 33, 107–120. doi: 10.1007/s10919-008-0065-7

Pell, M. D., Rothermich, K., Liu, P., Paulmann, S., Sethi, S., and Rigoulot, S. (2015). Preferential decoding of emotion from human non-linguistic vocalizations versus speech prosody. *Biol. Psychol.* 111, 14–25. doi: 10.1016/j.biopsycho.2015.08.008

Peng, G., and Chen, F. (2020). "Speech development in mandarin-speaking children," in *Speech Perception, Production and Acquisition: Multidisciplinary Approaches in Chinese Languages*. eds. H. Liu, F. Tsao and P. Li (United States: Springer)

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation (PhD thesis)*. Boston, MA, United States: Massachusetts Institute of Technology; Department of Linguistics and Philosophy.

Polka, L., and Bohn, O. S. (2011). Natural referent vowel (NRV) framework: an emerging view of early phonetic development. *J. Phon.* 39, 467–478. doi: 10.1016/j.wocn.2010.08.007

Prochnow, A., Erlandsson, S., Hesse, V., and Wermke, K. (2019). Does a 'musical' mother tongue influence cry melodies? A comparative study of Swedish and German newborns. *Music. Sci.* 23, 143–156. doi: 10.1177/1029864917733035

Quam, C., and Swingley, D. (2010). Phonological knowledge guides 2-year-olds' and adults' interpretation of salient pitch contours in word learning. *J. Mem. Lang.* 62, 135–150. doi: 10.1016/j.jml.2009.09.003

Quam, C., and Swingley, D. (2012). Development in children's interpretation of pitch cues to emotions. *Child Dev.* 83, 236–250. doi: 10.1111/j.1467-8624.2011.01700.x

Ramachers, S., Brouwer, S., and Fikkert, P. (2018). No perceptual reorganization for Limburgian tones? A cross-linguistic investigation with 6-to 12-month-old infants. *J. Child Lang.* 45, 290–318. doi: 10.1017/S0305000917000228

Reh, R. K., Dias, B. G., Nelson, C. A., Kaufer, D., Werker, J. F., Kolb, B., et al. (2020). Critical period regulation across multiple timescales. *Proc. Natl. Acad. Sci.* 117, 23242–23251. doi: 10.1073/pnas.1820836117

Reh, R. K., Hensch, T. K., and Werker, J. F. (2021). Distributional learning of speech sound categories is gated by sensitive periods. *Cognition* 213:104653. doi: 10.1016/j.cognition.2021.104653

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science* 212, 947–950. doi: 10.1126/science.7233191

Remijsen, B. (2002). "Lexically contrastive stress accent and lexical tone in Ma'ya," in *Laboratory Phonology 7* (Berlin: De Gruyter Mouton), 585–614.

Remijsen, B. (2016). "Tone," in *The Oxford Research Encyclopedia of Linguistics Online* (Oxford: Oxford University Press).

Ren, G. Q., Tang, Y. Y., Li, X. Q., and Sui, X. (2013). "Pre-attentive processing of mandarin tone and intonation: evidence from event-related potentials," in *Functional brain Mapping and the Endeavor to Understand the Working brain*. eds. F. Signorelli and D. Chirchiglia (Austria: Intech), 95–108.

Roseano, P., González, M., Borràs-Comes, J., and Prieto, P. (2016). Communicating epistemic stance: how speech and gesture patterns reflect epistemicity and evidentiality. *Discourse Process.* 53, 135–174. doi: 10.1080/0163853X.2014.969137

Ross, E. D., Edmondson, J. A., and Seibert, G. B. (1986). The effect of affect on various acoustic measures of prosody in tone and non-tone languages: a comparison based on computer analysis of voice. *J. Phon.* 14, 283–302. doi: 10.1016/S0095-4470(19)30669-2

Ruba, A. L., Meltzoff, A. N., and Repacholi, B. M. (2020). Superordinate categorization of negative facial expressions in infancy: The influence of labels. *Dev. Psychol.* 56, 671–685. doi: 10.1037/dev0000892

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928. doi: 10.1126/science.274.5294.1926

Saffran, J. R., and Kirkham, N. Z. (2018). Infant statistical learning. *Annu. Rev. Psychol.* 69, 181–203. doi: 10.1146/annurev-psych-122216-011805

Sambeth, A., Ruohio, K., Alku, P., Fellman, V., and Huotilainen, M. (2008). Sleeping newborns extract prosody from continuous speech. *Clin. Neurophysiol.* 119, 332–341. doi: 10.1016/j.clinph.2007.09.144

Sauter, D. A., Eisner, F., Ekman, P., and Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proc. Natl. Acad. Sci.* 107, 2408–2412. doi: 10.1073/pnas.0908239106

Sauter, D. A., Eisner, F., Ekman, P., and Scott, S. K. (2015). Emotional vocalizations are recognized across cultures regardless of the valence of distractors. *Psychol. Sci.* 26, 354–356. doi: 10.1177/0956797614560771

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychol. Bull.* 99, 143–165. doi: 10.1037/0033-2909.99.2.143

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Comm.* 40, 227–256. doi: 10.1016/S0167-6393(02)00084-5

Scott, S. K., and McGettigan, C. (2013). Do temporal processes underlie left hemisphere dominance in speech perception? *Brain Lang.* 127, 36–45. doi: 10.1016/j.bandl.2013.07.006

Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *J. Mem. Lang.* 57, 24–48. doi: 10.1016/j.jml.2006.10.004

Seidl, A., and Johnson, E. K. (2006). Infant word segmentation revisited: edge alignment facilitates target extraction. *Dev. Sci.* 9, 565–573. doi: 10.1111/j.1467-7687.2006.00534.x

Shaw, J. A., and Tyler, M. D. (2020). Effects of vowel coproduction on the timecourse of tone recognition. *J. Acoust. Soc. Am.* 147, 2511–2524. doi: 10.1121/10.0001103

Shi, R., Gao, J., Achim, A., and Li, A. (2017). Perception and representation of lexical tones in native mandarin-learning infants and toddlers. *Front. Psychol.* 8:1117. doi: 10.3389/fpsyg.2017.01117

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition* 106, 833–870. doi: 10.1016/j.cognition.2007.05.002

Singh, L., Fu, C. S., Seet, X. H., Tong, A. P., Wang, J. L., and Best, C. T. (2018). Developmental change in tone perception in mandarin monolingual, English monolingual, and mandarin–English bilingual infants: divergences between monolingual and bilingual learners. *J. Exp. Child Psychol.* 173, 59–77. doi: 10.1016/j.jecp.2018.03.012

Singh, L., Goh, H. H., and Wewalaarachchi, T. D. (2015). Spoken word recognition in early childhood: comparative effects of vowel, consonant and lexical tone variation. *Cognition* 142, 1–11. doi: 10.1016/j.cognition.2015.05.010

Singh, L., Hui, T. J., Chan, C., and Golinkoff, R. M. (2014). Influences of vowel and tone variation on emergent word knowledge: a cross-linguistic investigation. *Dev. Sci.* 17, 94–109. doi: 10.1111/desc.12097

Singh, L., Morgan, J. L., and Best, C. T. (2002). Infants' listening preferences: baby talk or happy talk? *Infancy* 3, 365–394. doi: 10.1207/S15327078IN0303_5

Singh, L., Morgan, J. L., and White, K. S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. *J. Mem. Lang.* 51, 173–189. doi: 10.1016/j.jml.2004.04.004

Singh, L., White, K. S., and Morgan, J. L. (2008). Building a word-form lexicon in the face of variable input: influences of pitch and amplitude on early spoken word recognition. *Lang. Learn. Dev.* 4, 157–178. doi: 10.1080/15475440801922131

Snow, D. (2001). Intonation in the monosyllabic utterances of 1-year-olds. *Infant Behav. Dev.* 24, 393–407. doi: 10.1016/S0163-6383(02)00084-X

Snow, D. (2006). Regression and reorganization of intonation between 6 and 23 months. *Child Dev.* 77, 281–296. doi: 10.1111/j.1467-8624.2006.00870.x

Snow, D., and Balog, H. L. (2002). Do children produce the melody before the words? A review of developmental intonation research. *Lingua* 112, 1025–1058. doi: 10.1016/S0024-3841(02)00060-8

So, C. K., and Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: effects of native phonological and phonetic influences. *Lang. Speech* 53, 273–293. doi: 10.1177/0023830909357156

So, C. K., and Best, C. T. (2011). Categorizing mandarin tones into listeners' native prosodic categories: The role of phonetic properties. *Poznań Stud. Contemp. Ling.* 47:133. doi: 10.2478/psicl-2011-0011

So, C. K., and Best, C. T. (2014). Phonetic influences on English and French listeners' assimilation of mandarin tones to native prosodic categories. *Stud. Second. Lang. Acquis.* 36, 195–221. doi: 10.1017/S0272263114000047

So, L. K., and Dodd, B. J. (1995). The acquisition of phonology by Cantonese-speaking children. *J. Child Lang.* 22, 473–495. doi: 10.1017/S0305000900009922

Soken, N. H., and Pick, A. D. (1999). Infants' perception of dynamic affective expressions: do infants distinguish specific expressions? *Child Dev.* 70, 1275–1282. doi: 10.1111/1467-8624.00093

Stern, D. N., Spieker, S., and MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. *Dev. Psychol.* 18, 727–735. doi: 10.1037/0012-1649.18.5.727

Sullivan, J. W., and Horowitz, F. D. (1983). The effects of intonation on infant attention: The role of the rising intonation contour. *J. Child Lang.* 10, 521–534. doi: 10.1017/S0305000900005341

Thiessen, E. D., Hill, E. A., and Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy* 7, 53–71. doi: 10.1207/s15327078in0701_5

Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., et al. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *J. Acoust. Soc. Am.* 137, 3005–3007. doi: 10.1121/1.4919349

To, C. K., Cheung, P. S., and McLeod, S. (2013). A population study of children's acquisition of Hong Kong Cantonese consonants, vowels, and tones. *J. Speech Lang. Hear. Res.* 56, 103–122. doi: 10.1044/1092-4388(2012/11-0080)

Tonkova-Yampol'skaya, R. V. (1969). Development of speech intonation in infants during the first two years of life. *Sov. Psychol.* 7, 48–54. doi: 10.2753/RPO1061-0405070348

Tsao, F. M. (2017). Perceptual improvement of lexical tones in infants: effects of tone language experience. *Front. Psychol.* 8:558. doi: 10.3389/fpsyg.2017.00558

Tyler, M. D., Best, C. T., Goldstein, L. M., and Antoniou, M. (2014). Investigating the role of articulatory organs and perceptual assimilation in infants' discrimination of native and non-native fricative place contrasts. *Dev. Psychobiol.* 56, 210–227. doi: 10.1002/dev.21195

van Heuven, V. J. (2018). Acoustic correlates and perceptual cues of word and sentence stress: towards a cross-linguistic perspective. In R. Goedemans, J. Heinz and HulstH. van der (Eds.), *The Study of word Stress and accent: Theories, Methods and data*. (15–59). England: Cambridge University Press.

Walker-Andrews, A. S., and Grolnick, W. (1983). Discrimination of vocal expressions by young infants. *Infant Behav. Dev.* 6, 491–498. doi: 10.1016/S0163-6383(83)90331-4

Walker-Andrews, A. S., and Lennon, E. (1991). Infants' discrimination of vocal expressions: contributions of auditory and visual information. *Infant Behav. Dev.* 14, 131–142. doi: 10.1016/0163-6383(91)90001-9

Wang, T., and Lee, Y. C. (2015). Does restriction of pitch variation affect the perception of vocal emotions in mandarin Chinese? *J. Acoust. Soc. Am.* 137, EL117–EL123. doi: 10.1121/1.4904916

Wang, T., Lee, Y. C., and Ma, Q. (2018). Within and across-language comparison of vocal emotions in mandarin and English. *Appl. Sci.* 8:2629. doi: 10.3390/app8122629

Wang, T., and Qian, Y. (2018). Are pitch variation cues indispensable to distinguish vocal emotions. In *Proceedings of the 9th International Conference on Speech Prosod*y 324–328.

Wanrooij, K., Boersma, P., and van Zuijen, T. L. (2014). Distributional vowel training is less effective for adults than for infants. A study using the mismatch response. *PLoS One* 9:e109806. doi: 10.1371/journal.pone.0109806

Wellman, H. M. (1992). *The child's Theory of mind*.United States: MIT Press.

Werker, J. F., and Hensch, T. K. (2015). Critical periods in speech perception: new directions. *Annu. Rev. Psychol.* 66, 173–196. doi: 10.1146/annurev-psych-010814-015104

Werker, J. F., and Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Percept. Psychophys.* 37, 35–44. doi: 10.3758/BF03207136

Wermke, K., Ruan, Y., Feng, Y., Dobnig, D., Stephan, S., Wermke, P., et al. (2017). Fundamental frequency variation in crying of mandarin and German neonates. *J. Voice* 31:255.e30. doi: 10.1016/j.jvoice.2016.06.009

Wermke, K., Teiser, J., Yovsi, E., Kohlenberg, P. J., Wermke, P., Robb, M., et al. (2016). Fundamental frequency variation within neonatal crying: does ambient

language matter? *Speech Lang. Hear.* 19, 211–217. doi: 10.1080/2050571X.2016.1187903

Wewalaarachchi, T. D., and Singh, L. (2016). Effects of suprasegmental phonological alternations on early word recognition: evidence from tone sandhi. *Front. Psychol.* 7:627. doi: 10.3389/fpsyg.2016.00627

Widen, S. C. (2013). Children's interpretation of facial expressions: The long path from valence-based to specific discrete categories. *Emot. Rev.* 5, 72–77. doi: 10.1177/1754073912451492

Wong, P. (2012a). Acoustic characteristics of three-year-olds' correct and incorrect monosyllabic mandarin lexical tone productions. *J. Phon.* 40, 141–151. doi: 10.1016/j.wocn.2011.10.005

Wong, P. (2012b). Monosyllabic mandarin tone productions by 3-year-olds growing up in Taiwan and in the United States: Interjudge reliability and perceptual results. *J. Speech Lang. Hear. Res.* 55, 1423–1437. doi: 10.1044/1092-4388(2012/11-0273)

Wong, P. (2013). Perceptual evidence for protracted development in monosyllabic mandarin lexical tone production in preschool children in Taiwan. *J. Acoust. Soc. Am.* 133, 434–443. doi: 10.1121/1.4768883

Wong, P., Schwartz, R. G., and Jenkins, J. J. (2005). Perception and production of lexical tones by 3-year-old, mandarin-speaking children. *J. Speech Lang. Hear. Res.* 48, 1065–1079. doi: 10.1044/1092-4388(2005/074)

Yanushevskaya, I., Gobl, C., and Ní Chasaide, A. (2018). Cross-language differences in how voice quality and f0 contours map to affect. *J. Acoust. Soc. Am.* 144, 2730–2750. doi: 10.1121/1.5066448

Yeung, H. H., Chen, K. H., and Werker, J. F. (2013). When does native language input affect phonetic perception? The precocious case of lexical tone. *J. Mem. Lang.* 68, 123–139. doi: 10.1016/j.jml.2012.09.004

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Yoshida, K. A., Pons, F., Maye, J., and Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. *Infancy* 15, 420–433. doi: 10.1111/j.1532-7078.2009.00024.x

Yuan, J. (2011). Perception of intonation in mandarin Chinese. *J. Acoust. Soc. Am.* 130, 4063–4069. doi: 10.1121/1.3651818

Zora, H., Schwarz, I. C., and Heldner, M. (2015). Neural correlates of lexical stress: mismatch negativity reflects fundamental frequency and intensity. *Neuroreport* 26, 791–796. doi: 10.1097/WNR.0000000000000426

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.