Evaluating relative perceptual salience of linguistic and emotional prosody in quiet and noisy contexts

Minyue Zhang , Hui Zhang , Enze Tang , Hongwei Ding * , Yang Zhang *

Posted Date: 8 September 2023

doi: 10.20944/preprints202309.0585.v1

Keywords: babble noise; lexical tone; emotional prosody; masking

Article

Evaluating Relative Perceptual Salience of Linguistic and Emotional Prosody in Quiet and Noisy Contexts

Minyue Zhang [1], Hui Zhang [1], Enze Tang [1], Hongwei Ding [1,*] and Yang Zhang [2,*]

[1] Speech-Language-Hearing Center, School of Foreign Languages, Shanghai Jiao Tong University, Shanghai 200240, China; zhang.my@sjtu.edu.cn (M.Z.); zhanghui_Helen@126.com (H.Z.); tangenze@sjtu.edu.cn (T.Z.)

[2] Department of Speech-Language-Hearing Sciences and Masonic Institute for the Developing Brain, University of Minnesota, Minneapolis, MN 55455, USA

* Correspondence: hwding@sjtu.edu.cn (H.D.); zhanglab@umn.edu (Y.Z.); Tel.: +1-612-624-7878 (Y.Z.)

Abstract: How people recognize linguistic and emotional prosody in different listening conditions is essential for understanding the complex interplay between social context, cognition, and communication. The perception of both lexical tones and emotional prosody depends on prosodic features including pitch, intensity, duration, and voice quality. However, it is unclear which aspect of prosody is perceptually more salient and resistant to noise. This study aimed to investigate the relative perceptual robustness of emotional prosody and lexical tone recognition in quiet and in the presence of multi-talker babble noise. Forty young adults with normal hearing listened to monosyllables either with or without background babble noise and completed two identification tasks, one for emotion recognition and the other for lexical tone recognition. Compared with emotional prosody, lexical tones were more perceptually salient in multi-talker babble noise. Native Mandarin Chinese participants identified lexical tones more accurately and quickly than vocal emotions at the same signal-to-noise ratio. Lexical tone perception is also more robust against babble speech noise degradation than emotional prosody perception for native Mandarin Chinese listeners. Acoustic and cognitive dissimilarities between linguistic prosody and emotional prosody may have led to the phenomenon, which calls for further explorations into the underlying psychobiological and neurophysiological mechanisms.

Keywords: babble noise; lexical tone; emotional prosody; masking

1. Introduction

In human communication, prosodic features of the spoken language fulfill important linguistic and socio-affective functions. Emotional prosody refers to the prosodic expression of the emotional state of the speaker [1], whereas linguistic prosody relates to the use of prosody to specify linguistic information [2]. While linguistic and emotional prosodies serve different communicative functions, both are acoustically characterized by variations in fundamental frequency (also referred to as F0 or pitch), intensity, duration, and voice quality [3–5]. Recognizing linguistic tone and emotional prosody is crucial for effective communication, as these cues provide information about the speaker's intent, mood, and the emotional content of their message.

In tonal languages such as Mandarin Chinese, pitch variations play a crucial role in distinguishing word meanings at the syllabic level, forming phonemic contrasts known as lexical tones [6]. Despite their importance for conveying phonological and semantic contrasts, lexical tones share some characteristics with prosody, such as their suprasegmental pitch variations and larynx-based articulation [7], and are therefore considered an important constituent of linguistic prosody [8]. Mandarin Chinese comprises four lexical tones differentiated by their pitch contours: high and flat (Tone 1), rising (Tone 2), falling and then rising (Tone 3), and falling (Tone 4). The perception of Mandarin lexical tones largely relies on fundamental frequency (F0) [9,10], with F0 contour and F0 height being the primary acoustic cues used to distinguish between the four tones [11–14]. Although the co-varying intensity and duration parameters in Mandarin speech provide supplementary/redundant perceptual cues [9,15], there is evidence that manipulating duration and amplitude may have little effect on lexical tone perception e.g., [16].

 doi:10.20944/preprints202309.0585.v1

Listening conditions play a significant role in how people perceive and interpret linguistic as well as emotional prosody. Everyday communication often takes place in noisy environments, such as bustling streets, crowded cafes, busy offices, or even during social events. These conditions can range from quiet environments with minimal background noise to noisy settings with various auditory distractions. In noisy contexts, individuals may encounter difficulties in accurately perceiving and distinguishing linguistic tone and emotional prosody due to reduced auditory clarity. This can lead to misinterpretations, misunderstandings, increased effort and cognitive load, and challenges in effective communication. The robustness of Mandarin lexical tone perception in adverse listening conditions has been well documented [17–22]. In the comparable signal-to-noise ratio (SNR) conditions for both steady-state and fluctuating maskers, Mandarin lexical tone recognition performances were found to be better than English sentence recognition [23]. Wang and Xu [22] further verified this phenomenon by observing that speech-shaped noise and multi-talker babble with various numbers of talkers had less impact on Mandarin lexical tone perception than on recognition of English vowel-consonant-vowel syllables, words, or sentences. The high robustness of lexical tones relative to other linguistic segmental elements (especially those in non-tonal languages) has been attributed to listeners' additional use of frequency-modulation information (referred to as temporal fine structure by Qi, et al. [21]) in tone perception. This feature is reported to be particularly resistant to background noise degradation [18,24–26].

Unlike lexical tones whose perception is highly related to the listener's linguistic knowledge and experience [27–29], emotional prosody conveys a broad range of emotional states, among which basic emotions (typically including happiness, sadness, anger, fear, disgust, and surprise [30]) can be recognized across cultures [31,32]. Basic emotional prosody displays a more universal feeling [33], and vocal emotion communication is constrained largely by biological factors [34] and governed by universal principles across languages and cultures [35,36]. However, these findings and views were primarily based on non-tonal languages. Later cross-linguistic comparisons have shown that despite the universality of emotional expressions, the specific mechanisms of utilizing acoustic cues for encoding emotions in various languages are still different e.g., [37]. Like lexical tones, acoustic parameters such as pitch, duration, and intensity have been found to be important for emotion identification [33,38–41]. Many studies additionally pointed out the significance of voice quality features in distinguishing emotions (e.g., anger and happiness [42,43]). In tonal languages, the existence of a lexical tone system may restrict the use of pitch for paralinguistic purposes [44], thus highlighting the importance of other acoustic cues, particularly voice quality, for conveying vocal emotions [37].

Most investigations into how background noise affects emotion recognition have focused on improving automatic emotion recognition using speech enhancement and artificial intelligence algorithms e.g., [45,46]. However, recent studies have started to explore how background noise influences emotion perception in human listeners e.g., [24,47–51]. For instance, Parada-Cabaleiro, et al. [48] investigated the effects of three types of background noise (white, pink, and Brownian) on emotional speech perception and found that all types of noise negatively impacted performance, with pink noise having the most significant effect and Brownian the least. Scharenborg, et al. [47] examined the influence of babble noise on verbal emotion perception in both native and foreign languages, while Zhang and Ding [49] explored how background babble noise affected emotion identification in unisensory and multisensory settings. The findings of these studies consistently demonstrate that background noise, particularly babble noise, can have detrimental effects on emotion perception.

Two theoretical accounts exist with opposing claims on the relative salience or functional weight of linguistic versus emotional prosody. According to the "functional load" hypothesis [52], lexical tones in tone languages carry a high functional load with phonemic status equivalent to that of vowels. Ross, et al. [53] extended this idea to examine emotional prosody in Mandarin Chinese, in comparison with English, and found that the use of tone in a language limits the extent to which F0 can be freely used to signal emotions. These findings suggest that linguistic prosody may be more salient than emotional prosody in tonal languages where tone is used to distinguish between different words. However, Xu [54] demonstrated that various aspects of prosody are encoded by different

 doi:10.20944/preprints202309.0585.v1

mechanisms that rely on F0 for different purposes, implying that tonal languages may not have a limited capacity for intonation for linguistic or paralinguistic functions. In contrast, the social signaling theory [34,55] posits that emotional prosody is crucial for nonverbal communication and conveys information about the speaker's emotional state, personality, social identity, intentions, and attitudes towards the listener. While both emotional prosody and linguistic prosody are important for social signaling, emotional prosody may be more salient because it communicates critical social and affective information.

While there is theoretical debate on the relative salience of linguistic and emotional prosody, few studies have empirically investigated their relative perceptual resilience under adverse listening conditions. Recent studies have shown that white noise has a greater impact on word recognition than emotional prosody recognition in English [24]. However, whether these results generalize to tonal languages such as Mandarin Chinese remains unclear. Moreover, previous studies have used different testing paradigms for assessing word/sentence recognition versus emotional prosody recognition (i.e., open-set tests for word/sentence recognition vs. forced-choice close-set tests for emotional prosody recognition), rendering the identification of emotions much simpler [21,22]. In addition, although white noise has been used in previous research, using multi-talker babble noise, which is commonly encountered in everyday listening environments [56,57], may provide a more ecologically valid measure of the impact of background noise on prosody perception. Researchers have observed that Mandarin lexical tone recognition remains robust even in adverse listening conditions, with performance plateauing at N = 8 in all SNR conditions when using multi-talker babble noise [22].

Given that everyday communication frequently occurs in noisy environments, understanding how people cope with these challenges and how they adapt their communication strategies is essential. The present study aimed to investigate the relative perceptual resilience of Mandarin lexical tones and emotional prosody in background multi-talker babble noise. We hypothesized that lexical tones would be more perceptually salient than emotional prosody under adverse listening conditions with masking babble noise. Understanding the relative salience of linguistic and emotional prosody in different listening conditions is crucial for ensuring effective communication and providing insights into improving communication strategies, enhancing educational experiences, and gaining a deeper understanding of human cognitive and emotional processes.

2. Methods

2.1. Participants

Forty young adults (21 females and 19 males, mean age ± SD: 22.19 ± 2.76 years old) were recruited to participate in the experiment through an online campus advertisement. All participants spoke Mandarin Chinese as their native language and the dominant language in daily use. All had normal hearing as verified by standard audiological screening for pure tones from 0.25–8 kHz (≤ 20dB HL) [58]. None reported a history of speech, language, hearing impairment, or any psychological or neurological conditions. All participants completed written informed consent before the experiment and were paid afterward for their participation.

2.2. Stimuli

Eight monosyllabic interjections, 嘿, 啊, 哎, 呀, 哈, 诶, 咳, and 哦 (International Phonetic Alphabet [xeɪ], [a], [aɪ], [ja], [xa], [eɪ], [xaɪ], and [ɔ]), were chosen to carry emotional prosody and lexical tones. We chose monosyllabic interjections out of two major considerations. One is that the carriers of emotional prosody and lexical tones should be the same to enable legitimate comparisons between them, and the other is about ecological validity. Interjections are important devices in conversations to express mental or emotional states [59], and monosyllables in Mandarin Chinese can be pronounced with one of the four lexical tones [60]. It is therefore ecologically valid to use monosyllabic interjections as the carriers in this experiment. Each monosyllable was produced with four emotions (happy, sad, angry, and calm) and four lexical tones (level tone, Tone 1; rising tone,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 September 2023 doi:10.20944/preprints202309.0585.v1

Tone 2; dipping tone, Tone 3; and falling tone, Tone 4) in a soundproof booth by two amateur actors (one female and one male) who were native speakers of Mandarin Chinese, resulting in 8 [interjections] × 8 [categories] × 2 [actors] = 128 [sound clips]. The sounds were recorded using a Neumann U87 Ai condenser studio microphone (Georg Neumann, Berlin, Germany) and a Fireface UFX soundboard (RME Fireface; RME Inc.) and were digitized at a sampling rate of 44,100 Hz with a 16-bit amplitude resolution, and normalized peak value (90%) using Adobe Audition CC (Adobe Systems, California). Thirty native Mandarin Chinese who did not take part in the current study were invited to validate the stimuli with the identification accuracy for each category being at least 90%.

The pitch, intensity, and duration measures of the prosodic stimuli are shown in Table 1. Pitch and intensity measurements were conducted on the vowel portion of the stimuli. The onset and offset of a pitch or intensity contour were determined by the beginning and cessation of periodicity of the waveform. For Tone 4 productions, since a substantial number of irregular cycles, indicating creakiness, was observed at the offset, the endpoint in such productions was determined by the last identifiable cycle. The contour was divided into 100 intervals of equal duration. F0 values in Hz and intensity values in dB were then obtained at the 101 time points and missing points in the middle caused by creakiness were interpolated using ProsodyPro [61] in Praat 6.0.37 [62]. The F0 and intensity values were manually checked for accuracy.

Table 1. Mean values (SD) of the acoustic measures for the prosodic stimuli: mean F0 (Hz), duration (msec), and mean intensity (dB).

Measure	Emotional prosody	Lexical tone
Mean F0 (Hz)	195.6 (76.1)	151.9 (46.3)
Duration (msec)	525.2 (185.6)	570.0 (84.6)
Mean intensity (dB)	77.7 (2.4)	78.0 (2.6)

The productions of lexical tone and emotional prosody stimuli were normalized using the T-value logarithmic transform to account for interspeaker variability in F0 range,

$$T = [(lgX - lgL)/(lgH - lgL)] \times 5, \quad (1)$$

where X represents the observed F0, and H and L are the maximum F0 and minimum F0, respectively, of the speaker [63]. Figure 1 displays the normalized pitch contours of the emotional prosody and lexical tone stimuli, averaged across all speakers and tokens. The pitch contours of the lexical tone stimuli adhere to the canonical contour of the four lexical tones in Mandarin Chinese [64] and the pitch contours of the emotional prosody stimuli closely resemble those reported by Li [65].

The stimuli were presented in two listening conditions (i.e., quiet and noise). For the noise condition, we used an eight-talker babble created by Chen, et al. [66] as the background noise. It was created by mixing eight emotionally neutral sentences produced by eight native Mandarin Chinese speakers. The babble noise was normalized peak value (90%) using Adobe Audition CC (Adobe Systems, California) and added to the target stimuli with the signal-to-noise ratio (SNR) being −13 dB, which was confirmed through the pilot testing to avoid ceiling performance yet partially mask the target sounds. The presentation of the babble noise began about 500 ms prior to the beginning of the target sound and ended about 500 ms following the target offset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 September 2023 doi:10.20944/preprints202309.0585.v1

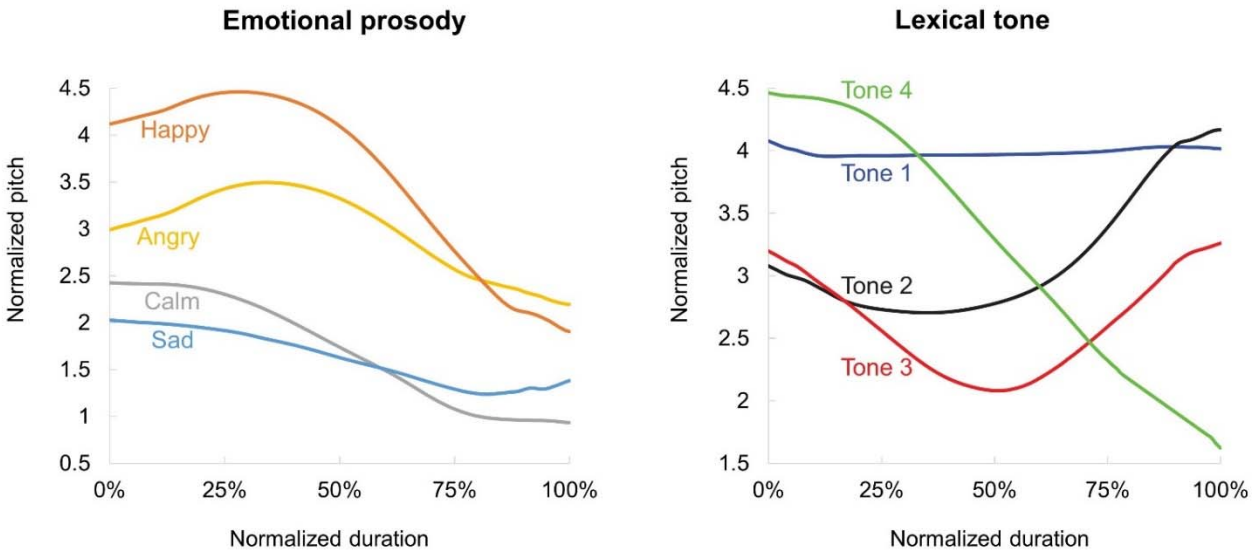

Figure 1. The pitch contours of the emotional prosody and the lexical tone stimuli. All prosodic contours were normalized to have the same duration, and the F0 values were log-transformed.

2.3. Procedure

The experiment was conducted in a sound-attenuating room with the participant seated at around 60 cm from an LCD monitor. We used Experiment Builder (Version 2.3.38; SR Research) for stimulus presentation. The sounds were presented binaurally using high-fidelity circumaural headphones (Sennheiser HD 280 Pro) at a comfortable level (70 dB SPL). There were two tasks, emotion recognition and tone recognition. In each task, the participants listened to a total of 128 stimuli (8 [interjections] × 4 [categories] × 2 [actors] × 2 [conditions]) that were presented in two blocks. Both the tasks and the blocks were counterbalanced across participants. Each block included 64 trials presented in a pseudorandom order. Participants were asked to respond as quickly and accurately as possible by pressing one of the four response keys that were mapped to the four emotional categories or the four lexical tones. Correspondences between the emotion/tone and the key were counterbalanced across participants but were held constant throughout the experiment for each participant. We carefully checked the participants' understanding of the general procedures as well as the correspondence between the keys and the categories before starting the experiment. Each block started with a practice phase consisting of four trials. Participants needed to reach 100% accuracy in the practice before entering the test phase. Breaks were inserted between blocks to avoid fatigue.

2.4. Statistical Analyses

To compare the masking effects of babble noise on emotional prosody and lexical tones, we applied a series of generalized linear mixed-effects models in R (Version 4.1.3) with the *lme4* package [67]. Accuracy and reaction time were entered as dependent variables respectively. For the analysis of accuracy, binomial response data were used and a binomial distribution with a logit link function was employed. For the analysis of reaction time, a gamma distribution with a log-link function was implemented [68]. Before analyzing reaction time data, we preprocessed them by excluding incorrect responses and responses over 2 SDs from the mean [69,70]. Within-subject variables, task (emotion and tone) and listening condition (quiet and noise) were entered as categorical fixed factors. Speakers and items were included as a random intercept term to account for the subject- and item-level variability. Tukey's post hoc tests in the *emmeans* package [71] were implemented for pairwise comparison when there was a significant main effect or interaction effect. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect

 doi:10.20944/preprints202309.0585.v1

in question. The full models with intercepts, coefficients, and error terms are respectively represented in formulas (1) and (2) in Supplemental Material S1.

3. Results

Supplemental Tables S1 and S2 summarize the detailed results of the generalized linear mixed-effects models for identification accuracy and reaction time.

3.1. Accuracy

Figure 2a illustrates the mean proportion correct in the quiet and noise listening conditions for the two tasks. Generalized linear mixed-effects analyses revealed significant main effects of task, $\chi^2(2) = 199.46$, $p < .001$, Cohen's $w = 2.23$, and condition, $\chi^2(2) = 1752.9$, $p < .001$, $w = 6.62$, and a significant interaction between task and condition, $\chi^2(1) = 27.75$, $p < .001$, $w = 0.83$. In the emotion recognition task, listeners achieved 35.9% ± 2.1% lower accuracy in the noise condition compared with the quiet condition ($\hat{\beta}_3 = 2.03$, $SE = .08$, $z = 26.45$, $p < .001$, $d = 2.39$). In the lexical tone recognition task, adding the same background babble noise led to a 29.9% ± 2.3% reduction in the identification accuracy ($\hat{\beta}_3 = 2.74$, $SE = .12$, $z = 23.42$, $p < .001$, $d = 3.23$). Lexical tone stimuli elicited 7.0% ± 0.9% more accurate responses than emotional prosody stimuli in the quiet condition ($\hat{\beta}_3 = -1.26$, $SE = .13$, $z = -9.97$, $p < .001$, $d = -1.48$), with the tone versus emotion gap further increased to 12.9% ± 1.4% in the noise condition ($\hat{\beta}_3 = -0.54$, $SE = .06$, $z = -9.16$, $p < .001$, $d = -.64$).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 September 2023 doi:10.20944/preprints202309.0585.v1

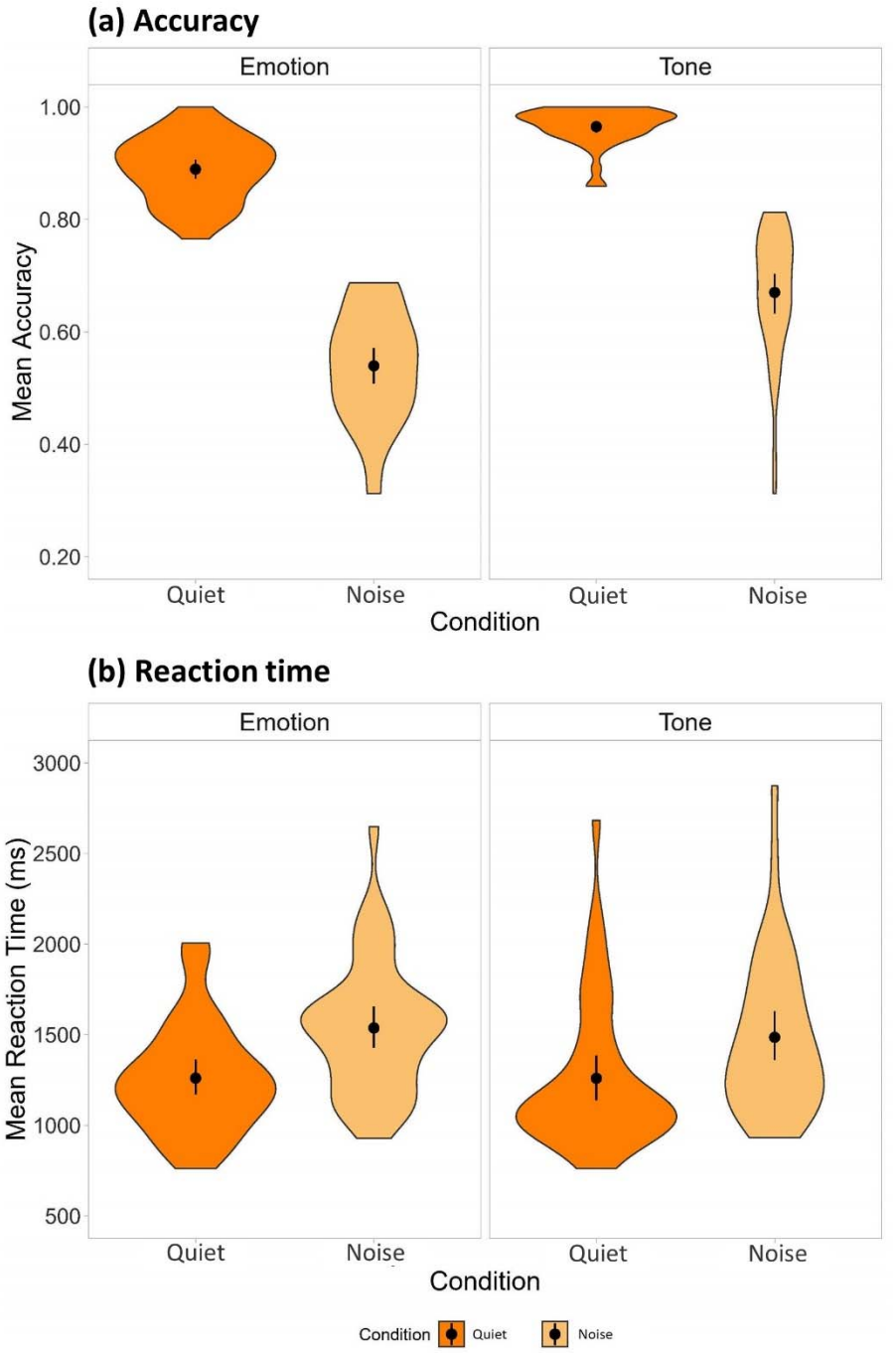

Figure 2. Mean (a) accuracy and (b) reaction time in the emotion and lexical tone recognition tasks. Mean accuracy and reaction time are displayed in the violin plots with data distribution shapes indicated by the density plots, mean values represented by the black dots, and 95% confidence intervals shown by the error bars.

3.2. Reaction time

For the reaction time data, we excluded incorrect responses (7.23% for the quiet condition and 39.90% for the noise condition) and responses over 2 SDs from the mean (5.16% for the quiet condition and 3.09% for the noise condition). Figure 2b illustrates the mean reaction time in the two listening conditions for the two tasks. Generalized linear mixed-effects analyses showed significant main effects of task, $\chi^2(2) = 20.16$, $p < .001$, $w = 0.71$, and condition, $\chi^2(2) = 408.73$, $p < .001$, $w = 3.20$, and a significant interaction between task and condition, $\chi^2(1) = 5.17$, $p = .02$, $w = 0.36$. In the emotion

 doi:10.20944/preprints202309.0585.v1

recognition task, response time was increased by 279.8 ± 21.8 ms in the noise condition compared with the quiet condition ($\hat{\beta}_3$ = −.20, *SE* = .013, *z* = -15.44, *p* < .001, *d* = −.27). Within the lexical tone recognition task, there was also a significant increase by 215.3 ± 18.6 ms in the noise condition relative to the quiet condition ($\hat{\beta}_3$ = −.16, *SE* = .012, *z* = −13.35, *p* < .001, *d* = −.22). Participants responded at 86.8 ± 21.0 ms faster to the lexical tone stimuli than to the emotional prosody stimuli in the noise condition ($\hat{\beta}_3$ = .06, *SE* = .014, *z* = 4.20, *p* < .001, *d* = .08), despite no significant difference between the two tasks in the quiet condition (*p* = .374).

4. Discussion

The current study investigated the relative perceptual resilience of Mandarin lexical tones and emotional prosody in background multi-talker babble noise. In line with our prediction, the accuracy and reaction time data showed a perceptual advantage of Mandarin lexical tones over emotional prosody. Specifically, native Mandarin Chinese speakers achieved higher identification accuracy and responded faster to the lexical tone stimuli, with these differences further amplified in the presence of masking babble noise. These findings align well with previous studies that have highlighted the robustness of Mandarin lexical tones to background noise e.g., [22]. Our results support the "functional load" account, which emphasizes the prominence of lexical tones over emotional prosody in tonal languages like Mandarin Chinese. We propose that the observed perceptual advantage of lexical tones can be attributed to both acoustic and cognitive differences between lexical tones and emotional prosody, as well as the specific characteristics of the masking babble noise used in this study.

Multi-talker babble noise produces two kinds of masking effects, that is, energetic masking (EM) and informational masking (IM). EM derives from the reduced audibility of the target because of the overlap in time and frequency between the signal and the masker, which is believed to influence processing from the level of the cochlea. IM arises from the similarity between the target and the masker despite the clear audibility of both and involves competition for resources in the central auditory system [72,73]. The mechanisms behind EM and IM can be explained through a framework based on auditory object formation and auditory object selection [74]. Object formation involves segregating the target source from maskers and object selection concerns selectively listening to the target while ignoring competing maskers. In our study, the eight-talker babble noise brought considerable difficulty in object formation with its high noise level but little in object selection due to its unintelligibility [75]. Hence, it brought about significant obstacles to extracting the acoustic features of the target stimuli but little lexical interference or competition for neural resources [76].

The acoustic characteristics of emotional prosody in Mandarin Chinese may have rendered its object formation more difficult in the presence of background noise. While the perception of Mandarin lexical tones depends majorly on pitch, the acoustic correlates of Mandarin emotional speech involve less contribution from pitch but more a crucial role of voice quality [77]. Since fundamental frequency is found to be more resistant to noise degradation than phonation-related cues [78,79], the extraction of acoustic cues for emotional prosody presumably would become harder than that for lexical tones in adverse listening conditions. Moreover, the acoustic realization of vocal emotions in Mandarin is characterized by its multidimensionality [37]. Due to the restricted paralinguistic use of pitch to accommodate the lexical tone system, other acoustic dimensions, including duration, intensity, and voice quality, are strengthened in compensation [37,80]. This may well increase the listeners' difficulty in integrating the necessary acoustic cues for emotion identification in the context of high-level background noise. Thus, the disadvantages in both extracting and integrating acoustic cues for emotional prosody together contributed to its less successful object formation in background noise. Admittedly, sources of difficulty could come from object selection – the other challenge of cocktail party listening. In our study, eight-talker babble noise introduced little linguistic interference because of its unintelligibility and thus might not have created a big obstacle for lexical tone perception. Rather, the speech elements in the masker could be competing for auditory attention, which would affect lexical tone recognition.

Another consideration is the psycho-cognitive differences between the two types of prosody. For each trial, listeners need to make cognitive evaluations of the target prosody [81] in attaching a label to the perceived prosodic expression. The cognitive evaluation process for emotional prosody might be less automatic than that for lexical tones because of the additional *conceptual* processing in the categorization of emotional expressions [82]. Numerous studies have documented a quite early acquisition and establishment of lexical tone categories [83,84] but not so for emotion perception. Emotional expressions are perceived in terms of valence in early development and become associated with discrete emotion categories over time as children learn emotion words [85]. It has been shown that the emotional specialization for vocal prosody occurs even later in adolescence [86]. Challenging listening environments may hinder the conceptual labelling for emotional prosody recognition and thus become especially disadvantageous to emotion perception.

Additionally, lexical tone recognition involves a strong top-down process [87–89] where prior language experience and linguistic knowledge promote the recognition of a pitch contour as a certain tone category [90]. As shown in Figure 1, the pitch contours of the lexical tone stimuli in this study exhibit a high degree of conformity to the canonical pitch contours of Mandarin Chinese lexical tones. The smaller reduction in the identification performances for lexical tones (as a type of linguistic prosody) thus aligns with the consensus view that top-down linguistic knowledge works well in compensating for the reduced informativeness of the bottom-up signals [91,92].

Both lower-level sensory and higher-level cognitive distinctions may be at work to influence the disparity of noise influences on the two types of prosody. That is, it might be more difficult to extract and integrate the acoustic cues of emotional prosody in babble noise due to its strong employment of noise-susceptible phonation-related parameters and its acoustic multidimensionality. It is also possible that the cognitive evaluation of emotional prosody before judgment involved additional conceptual processing that might be impeded in adverse conditions, whereas lexical tone recognition in noise may benefit from top-down facilitation driven by language experience, which can compensate for the signal loss from noise masking.

Our results are also consistent with the neurolinguistic view that prosody is processed in a hierarchical manner, that is, from sensory processing via auditory integration toward evaluative judgments [4,81,93]. This hierarchical 3-stage model of prosody perception may also be applicable in adverse listening environments. It remains unclear how emotional prosody and lexical tones resemble and differ from each other in terms of their neural underpinnings and mechanisms. In this regard, it is important to examine neural activations to determine at which stages of speech prosody perception involve more acoustic processing and at which stages the processing of functional classes (affective vs. linguistic) of speech prosody emerge. Do the two aspects happen discretely, or do they interact throughout the perception of prosodic information? Do emotional prosody and lexical tone perception in degraded conditions reflect the same functional hemispheric specialization as that in ideal listening environments? Answers to these questions may emerge when we disentangle the psychobiological and neurophysiological overlapping and non-overlapping between lexical tone processing and emotional prosody processing in both quiet and noise conditions.

There are limitations in this study. First, based on pilot testing, we chose only one specific SNR level for the noise condition to answer our hypothesis. It remains to be explored how variations in noise-induced degradation would affect the relative robustness of emotional prosody and lexical tones in background babble noise. Second, we chose only one type of noise (eight-talker babble) and did not incorporate other types of noise. Differences in the maskers may differentially impact lexical tone recognition and emotional prosody recognition. Third, communication involves more than just spoken words. Rather, it is a complex interplay of various sensory and modal cues that work together to convey meaning, emotions, and intentions. Our experimental protocol does not take into consideration the multimodal and multisensory nature of communication, which is essential for effective interpersonal interactions [5,94,95]. Speech communication is a holistic experience that involves integrating auditory, visual, tactile, and contextual cues to comprehend both the literal content and the emotional nuances of the message. This concept is particularly relevant in cross-cultural communication, where different cultures may rely on different modal cues to convey

 doi:10.20944/preprints202309.0585.v1

meaning and emotions, especially in adverse listening conditions. Moreover, this understanding has implications in fields like psychology, linguistics, and human-computer interaction, where researchers seek to create more realistic and natural communication models and technologies.

Our study provides an initial step for the comparison between the perception of emotional prosody and lexical tones in adverse listening conditions. Several lines can be pursued in the future. The first is to determine the role of language experience and linguistic knowledge in perceiving prosodic information in noise. Native tonal-language speakers may perform better in identifying linguistic prosody due to their tonal category knowledge. Different cultures may place varying degrees of emphasis on linguistic tone and emotional prosody [96–98]. Studying how these cues are interpreted across cultures and contexts can enhance intercultural communication and reduce misunderstandings. It would be enlightening to examine and compare the perception of emotional prosody and lexical tones in noise by non-tonal language speakers or Chinese-as-a-second-language learners in comparison with native speakers of Chinese. The second is to assess the relative masking effects of IM and EM on the two types of prosody by manipulating their proportion in background babble noise, which may be subject to influences of aging and aging-related hearing loss and cognitive decline [99–101]. The contribution of IM can be adjusted by varying the number of talkers in the babble noise or using speech samples from a non-tonal language unknown to the listeners to create babble noise. Speech-shaped noise can also be added for comparison purposes. The impact of noise on emotional prosody and lexical tones can depend on the type of noise and specific acoustic features of the speech signal. Babble or speech-shaped noise, for example, may have a greater effect on emotional prosody because they can disrupt the rhythm and timing of speech. Similarly, certain speech features such as pitch range or duration may be more critical for emotional prosody than for lexical tones, and therefore more susceptible to interference from noise. Furthermore, different SNR levels could be used to vary the degree of EM, which is typically greater at lower SNR levels [102]. Thirdly, it is important to consider how emotional prosody and lexical tones may interfere with each other [53,103,104]. Emotional prosody can make it harder to discern the subtle pitch differences that distinguish different lexical tones, while exaggerated or artificially manipulated lexical tones can alter the perception of emotional prosody. The extent of interference can depend on the specific task and context and may be symmetric or asymmetric. Individual differences in language proficiency, cognitive processing strategies, and attentional control can also affect the degree of interference. Additionally, the role of vowels/syllables may also need to be taken into consideration in this interaction. Finally, utilizing neurophysiological and neuroimaging techniques such as ERP and fMRI to record neural activity during the processing of emotional prosody and lexical tones in noise would help capture acoustic, psychobiological, and neurofunctional similarities and differences between various categories of prosodic information [7,105–109]. This approach can provide valuable insights into how the brain processes and distinguishes between different types of prosody, which have implications for individuals with perception/production difficulties with speech prosody [110–114].

5. Conclusion

Given that everyday communication frequently occurs in noisy environments, understanding how people cope with these challenges and how they adapt their communication strategies is essential. This study investigated the perception of Mandarin lexical tones and emotional prosody in quiet and in background multi-talker babble noise. Compared with emotional prosody, Mandarin lexical tones were more robust, less susceptible to background noise, and were more perceptually salient in noise. The stronger resilience of lexical tones in babble noise is in line with the distinctions between the two types of prosody at the three stages of the hierarchical model for prosody perception, which provides the impetus for further exploring the neural substrates of emotional prosody perception and lexical tone perception as well as their temporal and regional overlapping. By investigating the relative salience of linguistic and emotional prosody, researchers can provide insights into improving communication strategies in various populations who have difficulties with prosody processing, enhancing educational experiences, and gaining a deeper understanding of human cognitive and emotional processes.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Table S1: Generalized linear mixed-effects model with task and condition as the fixed effects, and accuracy as the dependent variable (pairwise contrasts are indented); Table S2: Generalized linear mixed-effects model with task and condition as the fixed effects, and reaction time as the dependent variable (pairwise contrasts are indented).

Author Contributions: Conceptualization, M.Z., H.D. and Y.Z.; methodology, M.Z., H.D. and Y.Z.; software, M.Z.; validation, H.D.; formal analysis, M.Z.; investigation, M.Z.; resources, M.Z., H.Z. and E.T.; data curation, M.Z.; writing—original draft preparation, M.Z.; writing—review and editing, H.D. and Y.Z.; visualization, M.Z.; supervision, H.D.; project administration, H.D. and Y.Z.; funding acquisition, H.D. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by grants from the Major Project of National Social Science Foundation of China (18ZDA293). Y. Zhang received additional support from University of Minnesota's Grand Challenges Exploratory Research Grant.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of School of Foreign Languages, Shanghai Jiao Tong University (2111S1218; Date of Approval: 8 December 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://osf.io/r8nmk/?view_only=6ad5e69885ba48cea3ab69ee77ed84a8.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Fairbanks, G.; Pronovost, W. Vocal Pitch During Simulated Emotion. *Science* **1938**, *88*, 382-383, doi:10.1126/science.88.2286.382.
2. Monrad-Krohn, G.H. The Prosodic Quality of Speech and Its Disorders: (a Brief Survey from a Neurologist's Point of View). *Acta Psychiatr. Scand.* **1947**, *22*, 255-269, doi:10.1111/j.1600-0447.1947.tb08246.x.
3. Cutler, A.; Pearson, M. On the Analysis of Prosodic Turn-Taking Cues. In *Intonation in Discourse*, Johns-Lewis, C., Ed.; Routledge: Abingdon, UK, 2018; pp. 139–155.
4. Belyk, M.; Brown, S. Perception of Affective and Linguistic Prosody: An ALE Meta-Analysis of Neuroimaging Studies. *Soc. Cogn. Affect. Neurosci.* **2014**, *9*, 1395-1403, doi:10.1093/scan/nst124.
5. Ding, H.; Zhang, Y. Speech Prosody in Mental Disorders. *Annu. Rev. Linguist.* **2023**, *9*, doi:10.1146/annurev-linguistics-030421-065139.
6. Fromkin, V.A.; Curtiss, S.; Hayes, B.P.; Hyams, N.; Keating, P.A.; Koopman, H.; Steriade, D. *Linguistics: An Introduction to Linguistic Theory*; Blackwell Oxford: Oxford, 2000.
7. Liang, B.; Du, Y. The Functional Neuroanatomy of Lexical Tone Perception: An Activation Likelihood Estimation Meta-Analysis. *Front. Neurosci.* **2018**, *12*, 495, doi:10.3389/fnins.2018.00495.
8. Gandour, J.T. Neural Substrates Underlying the Perception of Linguistic Prosody. In *Volume 2 Experimental Studies in Word and Sentence Prosody*, Carlos, G., Tomas, R., Eds.; De Gruyter Mouton: Berlin, New York, 2009; pp. 3-26.
9. Massaro, D.W.; Cohen, M.M.; Tseng, C.-y. The Evaluation and Integration of Pitch Height and Pitch Contour in Lexical Tone Perception in Mandarin Chinese. *Journal of Chinese Linguistics* **1985**, *13*, 267-289.
10. Xu, Y. Contextual Tonal Variations in Mandarin. *J. Phon.* **1997**, *25*, 61-83, doi:10.1006/jpho.1996.0034.
11. Jongman, A.; Wang, Y.; Moore, C.B.; Sereno, J.A. Perception and Production of Mandarin Chinese Tones. In *The Handbook of East Asian Psycholinguistics: Volume 1: Chinese*, Bates, E., Tan, L.H., Tzeng, O.J.L., Li, P., Eds.; Cambridge University Press: Cambridge, 2006; Volume 1, pp. 209-217.
12. Moore, C.B.; Jongman, A. Speaker Normalization in the Perception of Mandarin Chinese Tones. *J. Acoust. Soc. Am.* **1997**, *102*, 1864-1877, doi:10.1121/1.420092.
13. Yu, K.; Zhou, Y.; Li, L.; Su, J.a.; Wang, R.; Li, P. The Interaction between Phonological Information and Pitch Type at Pre-Attentive Stage: An ERP Study of Lexical Tones. *Lang. Cogn. Neurosci.* **2017**, *32*, 1164-1175, doi:10.1080/23273798.2017.1310909.

14. Tseng, C.-y.; Massaro, D.W.; Cohen, M.M. Lexical Tone Perception in Mandarin Chinese: Evaluation and Integration of Acoustic Features. In *Linguistics, Psychology, and the Chinese Language*, Kao, H.S.R., Hoosain, R., Eds.; Centre of Asian Studies, University of Hong Kong: 1986; pp. 91-104.
15. Whalen, D.H.; Xu, Y. Information for Mandarin Tones in the Amplitude Contour and in Brief Segments. *Phonetica* **1992**, *49*, 25-47, doi:10.1159/000261901.
16. Lin, M.-C. The Acoustic Characteristics and Perceptual Cues of Tones in Standard Chinese [Putonghua Shengdiao De Shengxue Texing He Zhijue Zhengzhao]. *Chinese Linguistics [Zhongguo Yuwen]* **1988**, *204*, 182-193.
17. Xu, L.; Tsai, Y.; Pfingst, B.E. Features of Stimulation Affecting Tonal-Speech Perception: Implications for Cochlear Prostheses. *J. Acoust. Soc. Am.* **2002**, *112*, 247-258, doi:10.1121/1.1487843.
18. Kong, Y.-Y.; Zeng, F.-G. Temporal and Spectral Cues in Mandarin Tone Recognition. *J. Acoust. Soc. Am.* **2006**, *120*, 2830-2840, doi:10.1121/1.2346009.
19. Krenmayr, A.; Qi, B.; Liu, B.; Liu, H.; Chen, X.; Han, D.; Schatzer, R.; Zierhofer, C.M. Development of a Mandarin Tone Identification Test: Sensitivity Index *D'* as a Performance Measure for Individual Tones. *Int. J. Audiol.* **2011**, *50*, 155-163, doi:10.3109/14992027.2010.530613.
20. Lee, C.-Y.; Tao, L.; Bond, Z.S. Effects of Speaker Variability and Noise on Mandarin Tone Identification by Native and Non-Native Listeners. *Speech Lang. Hear.* **2013**, *16*, 46-54, doi:10.1179/2050571X12Z.0000000003.
21. Qi, B.; Mao, Y.; Liu, J.; Liu, B.; Xu, L. Relative Contributions of Acoustic Temporal Fine Structure and Envelope Cues for Lexical Tone Perception in Noise. *J. Acoust. Soc. Am.* **2017**, *141*, 3022-3029, doi:10.1121/1.4982247.
22. Wang, X.; Xu, L. Mandarin Tone Perception in Multiple-Talker Babbles and Speech-Shaped Noise. *J. Acoust. Soc. Am.* **2020**, *147*, EL307-EL313, doi:10.1121/10.0001002.
23. Apoux, F.; Yoho, S.E.; Youngdahl, C.L.; Healy, E.W. Role and Relative Contribution of Temporal Envelope and Fine Structure Cues in Sentence Recognition by Normal-Hearing Listeners. *J. Acoust. Soc. Am.* **2013**, *134*, 2205-2212, doi:10.1121/1.4816413.
24. Morgan, S.D. Comparing Emotion Recognition and Word Recognition in Background Noise. *J. Speech Lang. Hear. Res.* **2021**, *64*, 1758-1772, doi:10.1044/2021_JSLHR-20-00153.
25. Lakshminarayanan, K.; Ben Shalom, D.; van Wassenhove, V.; Orbelo, D.; Houde, J.; Poeppel, D. The Effect of Spectral Manipulations on the Identification of Affective and Linguistic Prosody. *Brain Lang.* **2003**, *84*, 250-263, doi:10.1016/S0093-934X(02)00516-3.
26. van Zyl, M.; Hanekom, J.J. Speech Perception in Noise: A Comparison between Sentence and Prosody Recognition. *J. Hear. Sci.* **2011**, *1*, 54-56.
27. Krishnan, A.; Xu, Y.; Gandour, J.; Cariani, P. Encoding of Pitch in the Human Brainstem Is Sensitive to Language Experience. *Cogn. Brain Res.* **2005**, *25*, 161-168, doi:10.1016/j.cogbrainres.2005.05.004.
28. Klein, D.; Zatorre, R.J.; Milner, B.; Zhao, V. A Cross-Linguistic PET Study of Tone Perception in Mandarin Chinese and English Speakers. *NeuroImage* **2001**, *13*, 646-653, doi:10.1006/nimg.2000.0738.
29. Braun, B.; Johnson, E.K. Question or Tone 2? How Language Experience and Linguistic Function Guide Pitch Processing. *J. Phon.* **2011**, *39*, 585-594, doi:10.1016/j.wocn.2011.06.002.
30. Ekman, P. An Argument for Basic Emotions. *Cogn. Emot.* **1992**, *6*, 169–200, doi:10.1080/02699939208411068.
31. Sauter, D.A.; Eisner, F.; Ekman, P.; Scott, S.K. Cross-Cultural Recognition of Basic Emotions through Nonverbal Emotional Vocalizations. *Proc. Natl. Acad. Sci* **2010**, *107*, 2408-2412, doi:10.1073/pnas.0908239106.
32. Scherer, K.R.; Banse, R.; Wallbott, H.G. Emotion Inferences from Vocal Expression Correlate across Languages and Cultures. *J. Cross Cult. Psychol.* **2001**, *32*, 76-92, doi:10.1177/0022022101032001009.
33. Pell, M.D.; Monetta, L.; Paulmann, S.; Kotz, S.A. Recognizing Emotions in a Foreign Language. *J. Nonverbal Behav.* **2009**, *33*, 107-120, doi:10.1007/s10919-008-0065-7.
34. Scherer, K.R. Vocal Affect Expression: A Review and a Model for Future Research. *Psychol. Bull.* **1986**, *99*, 143-165, doi:10.1037/0033-2909.99.2.143.
35. Liu, P.; Pell, M.D. Processing Emotional Prosody in Mandarin Chinese: A Cross-Language Comparison. In *Proceedings of the International Conference on Speech Prosody*; 2014; pp. 95-99.
36. Bryant, G.A.; Barrett, H.C. Vocal Emotion Recognition across Disparate Cultures. *J. Cogn. Cult.* **2008**, *8*, 135-148, doi:10.1163/156770908X289242.
37. Wang, T.; Lee, Y.-c.; Ma, Q. Within and across-Language Comparison of Vocal Emotions in Mandarin and English. *Applied Sciences* **2018**, *8*, 2629.

38. Banse, R.; Scherer, K.R. Acoustic Profiles in Vocal Emotion Expression. *J. Pers. Soc. Psychol.* **1996**, *70*, 614-636, doi:10.1037/0022-3514.70.3.614.
39. Dupuis, K.; Pichora-Fuller, M.K. Intelligibility of Emotional Speech in Younger and Older Adults. *Ear Hear.* **2014**, *35*, 695-707, doi:10.1097/aud.0000000000000082.
40. Castro, S.L.; Lima, C.F. Recognizing Emotions in Spoken Language: A Validated Set of Portuguese Sentences and Pseudosentences for Research on Emotional Prosody. *Behav. Res. Methods* **2010**, *42*, 74-81, doi:10.3758/BRM.42.1.74.
41. Murray, I.R.; Arnott, J.L. Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. *J. Acoust. Soc. Am.* **1993**, *93*, 1097-1108, doi:10.1121/1.405558.
42. Juslin, P.N.; Laukka, P. Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code? *Psychol. Bull.* **2003**, *129*, 770-814, doi:10.1037/0033-2909.129.5.770.
43. Zhang, S. Emotion Recognition in Chinese Natural Speech by Combining Prosody and Voice Quality Features. Sun, F., Zhang, J., Tan, Y., Cao, J., Yu, W., Eds.; Advances in Neural Networks; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp. 457-464.
44. Hirst, D.; Wakefield, J.; Li, H.Y. Does Lexical Tone Restrict the Paralinguistic Use of Pitch? Comparing Melody Metrics for English, French, Mandarin and Cantonese. In *Proceedings of the International Conference on the Phonetics of Languages in China*; 2013; pp. 15–18.
45. Zhao, X.; Zhang, S.; Lei, B. Robust Emotion Recognition in Noisy Speech Via Sparse Representation. *Neural Comput. Appl.* **2014**, *24*, 1539-1553, doi:10.1007/s00521-013-1377-z.
46. Schuller, B.; Arsic, D.; Wallhoff, F.; Rigoll, G.I., Dresden. Emotion Recognition in the Noise Applying Large Acoustic Feature Sets. In *Proceedings of Speech Prosody 2006*; 2006; p. paper 128.
47. Scharenborg, O.; Kakouros, S.; Koemans, J. The Effect of Noise on Emotion Perception in an Unknown Language. In *Proceedings of the International Conference on Speech Prosody*; 2018; pp. 364–368.
48. Parada-Cabaleiro, E.; Baird, A.; Batliner, A.; Cummins, N.; Hantke, S.; Schuller, B. The Perception of Emotions in Noisified Non-Sense Speech. In *Proceedings of Interspeech 2017*; 2017; pp. 3246–3250.
49. Zhang, M.; Ding, H. Impact of Background Noise and Contribution of Visual Information in Emotion Identification by Native Mandarin Speakers. In *Proceedings of Interspeech 2022*; 2022; pp. 1993-1997.
50. Parada-Cabaleiro, E.; Batliner, A.; Baird, A.; Schuller, B. The Perception of Emotional Cues by Children in Artificial Background Noise. *Int. J. Speech Technol.* **2020**, *23*, 169-182, doi:10.1007/s10772-020-09675-1.
51. Luo, X. Talker Variability Effects on Vocal Emotion Recognition in Acoustic and Simulated Electric Hearing. *J. Acoust. Soc. Am.* **2016**, *140*, EL497-EL503, doi:10.1121/1.4971758.
52. Hockett, C. The Quantification of Functional Load. *Word* **1967**, *23*, 320–339, doi:10.1080/00437956.1967.11435484.
53. Ross, E.D.; Edmondson, J.A.; Seibert, G.B. 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.* **1986**, *14*, 283-302, doi:10.1016/S0095-4470(19)30669-2.
54. Xu, Y. Prosody, Tone and Intonation. In *The Routledge Handbook of Phonetics*, Katz, W.F., Assmann, P.F., Eds.; Routledge: New York, 2019; pp. 314-356.
55. Scherer, K.R.; Wallbott, H.G. Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning. *J. Pers. Soc. Psychol.* **1994**, *66*, 310–328, doi:10.1037//0022-3514.66.2.310.
56. Viswanathan, V.; Shinn-Cunningham, B.G.; Heinz, M.G. Temporal Fine Structure Influences Voicing Confusions for Consonant Identification in Multi-Talker Babble. *J. Acoust. Soc. Am.* **2021**, *150*, 2664-2676, doi:10.1121/10.0006527.
57. Cherry, E.C. Some Experiments on the Recognition of Speech, with One and with Two Ears. *J. Acoust. Soc. Am.* **1953**, *25*, 975-979, doi:10.1121/1.1907229.
58. Koerner, T.K.; Zhang, Y. Differential Effects of Hearing Impairment and Age on Electrophysiological and Behavioral Measures of Speech in Noise. *Hear. Res.* **2018**, *370*, 130-142, doi:10.1016/j.heares.2018.10.009.
59. Ameka, F. Interjections: The Universal yet Neglected Part of Speech. *J. Pragmat.* **1992**, *18*, 101-118, doi:10.1016/0378-2166(92)90048-G.
60. Howie, J.M. On the Domain of Tone in Mandarin. *Phonetica* **1974**, *30*, 129-148, doi:10.1159/000259484.
61. Xu, Y. Prosodypro—a Tool for Large-Scale Systematic Prosody Analysis. In *Tools and Resources for the Analysis of Speech Prosody*; Laboratoire Parole et Langage: Aix-en-Provence, France, 2013; pp. 7-10.
62. Boersma, P.; Weenink, D. *Praat: Doing Phonetics by Computer*, 6.0.37; 2018.

63. Wang, Y.; Jongman, A.; Sereno, J.A. Acoustic and Perceptual Evaluation of Mandarin Tone Productions before and after Perceptual Training. *J. Acoust. Soc. Am.* **2003**, *113*, 1033-1043, doi:10.1121/1.1531176.
64. Liu, S.; Samuel, A.G. Perception of Mandarin Lexical Tones When F0 Information Is Neutralized. *Lang. Speech* **2004**, *47*, 109-138, doi:10.1177/00238309040470020101.
65. Li, A. Emotional Intonation and Its Boundary Tones in Chinese. In *Encoding and Decoding of Emotional Speech: A Cross-Cultural and Multimodal Study between Chinese and Japanese*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2015; pp. 133-164.
66. Chen, F.; Hu, Y.; Yuan, M. Evaluation of Noise Reduction Methods for Sentence Recognition by Mandarin-Speaking Cochlear Implant Listeners. *Ear Hear.* **2015**, *36*, doi:10.1097/AUD.0000000000000074.
67. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using Lme4. *J. Stat. Softw.* **2015**, *67*, 1–48, doi:10.18637/jss.v067.i01.
68. Lo, S.; Andrews, S. To Transform or Not to Transform: Using Generalized Linear Mixed Models to Analyse Reaction Time Data. *Front. Psychol.* **2015**, *6*, 1171, doi:10.3389/fpsyg.2015.01171.
69. Baayen, R.H.; Milin, P. Analyzing Reaction Times. *Int. J. Psychol. Res.* **2010**, *3*, 12–28, doi:10.21500/20112084.807.
70. Chien, Y.F.; Sereno, J.A.; Zhang, J. What's in a Word: Observing the Contribution of Underlying and Surface Representations. *Lang. Speech* **2017**, *60*, 643–657, doi:10.1177/0023830917690419.
71. Lenth, R. *Emmeans: Estimated Marginal Means, Aka Leastsquares Means*, R package: 2020.
72. Brungart, D.S.; Simpson, B.D.; Ericson, M.A.; Scott, K.R. Informational and Energetic Masking Effects in the Perception of Multiple Simultaneous Talkers. *J. Acoust. Soc. Am.* **2001**, *110*, 2527-2538, doi:10.1121/1.1408946.
73. Scott, S.K.; McGettigan, C. The Neural Processing of Masked Speech. *Hear. Res.* **2013**, *303*, 58-66, doi:10.1016/j.heares.2013.05.001.
74. Shinn-Cunningham, B.G. Object-Based Auditory and Visual Attention. *Trends Cogn. Sci.* **2008**, *12*, 182-186, doi:10.1016/j.tics.2008.02.003.
75. Mattys, S.L.; Brooks, J.; Cooke, M. Recognizing Speech under a Processing Load: Dissociating Energetic from Informational Factors. *Cogn. Psychol.* **2009**, *59*, 203-243, doi:10.1016/j.cogpsych.2009.04.001.
76. Rosen, S.; Souza, P.; Ekelund, C.; Majeed, A. Listening to Speech in a Background of Other Talkers: Effects of Talker Number and Noise Vocoding. *J. Acoust. Soc. Am.* **2013**, *133*, 2431-2443, doi:10.1121/1.4794379.
77. Wang, T.; Ding, H.; Kuang, J.; Ma, Q. Mapping Emotions into Acoustic Space: The Role of Voice Quality. In *Proceedings of Interspeech 2014*; 2014; pp. 1978-1982.
78. Ingrisano, D.R.-S.; Perry, C.K.; Jepson, K.R. Environmental Noise. *Am. J. Speech Lang. Pathol.* **1998**, *7*, 91-96, doi:doi:10.1044/1058-0360.0701.91.
79. Perry, C.K.; Ingrisano, D.R.S.; Palmer, M.A.; McDonald, E.J. Effects of Environmental Noise on Computer-Derived Voice Estimates from Female Speakers. *J. Voice* **2000**, *14*, 146-153, doi:10.1016/S0892-1997(00)80021-1.
80. Wang, T.; Lee, Y.-c. Does Restriction of Pitch Variation Affect the Perception of Vocal Emotions in Mandarin Chinese? *J. Acoust. Soc. Am.* **2015**, *137*, EL117, doi:10.1121/1.4904916.
81. Schirmer, A.; Kotz, S.A. Beyond the Right Hemisphere: Brain Mechanisms Mediating Vocal Emotional Processing. *Trends Cogn. Sci.* **2006**, *10*, 24-30, doi:10.1016/j.tics.2005.11.009.
82. Fugate, J.M.B. Categorical Perception for Emotional Faces. *Emot. Rev.* **2013**, *5*, 84-89, doi:10.1177/1754073912451350.
83. Singh, L.; Fu, C.S.L. A New View of Language Development: The Acquisition of Lexical Tone. *Child Dev.* **2016**, *87*, 834-854, doi:10.1111/cdev.12512.
84. Yeung, H.H.; Chen, K.H.; Werker, J.F. When Does Native Language Input Affect Phonetic Perception? The Precocious Case of Lexical Tone. *J. Mem. Lang.* **2013**, *68*, 123-139, doi:10.1016/j.jml.2012.09.004.
85. Shablack, H.; Lindquist, K.A. The Role of Language in Emotional Development. In *Handbook of Emotional Development*, LoBue, V., Pérez-Edgar, K., Buss, K.A., Eds.; Springer International Publishing: Cham, 2019; pp. 451-478.
86. Morningstar, M.; Venticinque, J.; Nelson, E.E. Differences in Adult and Adolescent Listeners' Ratings of Valence and Arousal in Emotional Prosody. *Cogn. Emot.* **2019**, *33*, 1497-1504, doi:10.1080/02699931.2018.1561422.
87. Zhao, L.; Sloggett, S.; Chodroff, E. Top-Down and Bottom-up Processing of Familiar and Unfamiliar Mandarin Dialect Tone Systems. In *Proceedings of Speech Prosody 2022*; 2022; pp. 842-846.

88. Zhao, T.C.; Kuhl, P.K. Top-Down Linguistic Categories Dominate over Bottom-up Acoustics in Lexical Tone Processing. *J. Acoust. Soc. Am.* **2015**, *137*, 2379-2379, doi:10.1121/1.4920645.
89. Malins, J.G.; Gao, D.; Tao, R.; Booth, J.R.; Shu, H.; Joanisse, M.F.; Liu, L.; Desroches, A.S. Developmental Differences in the Influence of Phonological Similarity on Spoken Word Processing in Mandarin Chinese. *Brain Lang.* **2014**, *138*, 38-50, doi:10.1016/j.bandl.2014.09.002.
90. Shuai, L.; Gong, T. Temporal Relation between Top-Down and Bottom-up Processing in Lexical Tone Perception. *Front. Behav. Neurosci.* **2014**, *8*, 97, doi:10.3389/fnbeh.2014.00097.
91. Başkent, D. Effect of Speech Degradation on Top-Down Repair: Phonemic Restoration with Simulations of Cochlear Implants and Combined Electric–Acoustic Stimulation. *J. Assoc. Res. Otolaryngol.* **2012**, *13*, 683-692, doi:10.1007/s10162-012-0334-3.
92. Wang, J.; Shu, H.; Zhang, L.; Liu, Z.; Zhang, Y. The Roles of Fundamental Frequency Contours and Sentence Context in Mandarin Chinese Speech Intelligibility. *J. Acoust. Soc. Am.* **2013**, *134*, EL91-EL97.
93. Sammler, D.; Grosbras, M.-H.; Anwander, A.; Bestelmeyer, Patricia E.G.; Belin, P. Dorsal and Ventral Pathways for Prosody. *Curr. Biol.* **2015**, *25*, 3079-3085, doi:https://doi.org/10.1016/j.cub.2015.10.009.
94. Coulson, S. Sensorimotor Account of Multimodal Prosody. *PsyArXiv* **2023**, doi:10.31234/osf.io/2sc68.
95. Holler, J.; Levinson, S.C. Multimodal Language Processing in Human Communication. *Trends Cogn. Sci.* **2019**, *23*, 639-652.
96. Bryant, G.A. Vocal Communication across Cultures: Theoretical and Methodological Issues. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2022**, *377*, 20200387.
97. Lecumberri, M.L.G.; Cooke, M.; Cutler, A. Non-Native Speech Perception in Adverse Conditions: A Review. *Speech Commun.* **2010**, *52*, 864-886.
98. Liu, P.; Rigoulot, S.; Pell, M.D., , . https://doi.org/10.3389/fnhum.2015.00311. Cultural Differences in on-Line Sensitivity to Emotional Voices: Comparing East and West. *Front. Hum. Neurosci.* **2015**, *9*, 311, doi:10.3389/fnhum.2015.00311.
99. Gordon-Salant, S.; Fitzgibbons, P.J. Effects of Stimulus and Noise Rate Variability on Speech Perception by Younger and Older Adults. *J. Acoust. Soc. Am.* **2004**, *115*, 1808-1817.
100. Goossens, T.; Vercammen, C.; Wouters, J.; van Wieringen, A. Masked Speech Perception across the Adult Lifespan: Impact of Age and Hearing Impairment. *Hear. Res.* **2017**, *344*, 109-124.
101. Van Engen, K.J.; Phelps, J.E.; Smiljanic, R.; Chandrasekaran, B. Enhancing Speech Intelligibility: Interactions among Context, Modality, Speech Style, and Masker. *J. Speech Lang. Hear. Res.* **2014**, *57*, 1908-1918.
102. Scott, S.K.; Rosen, S.; Wickham, L.; Wise, R.J.S. A Positron Emission Tomography Study of the Neural Basis of Informational and Energetic Masking Effects in Speech Perception. *J. Acoust. Soc. Am.* **2004**, *115*, 813-821, doi:10.1121/1.1639336.
103. Nygaard, L.C.; Queen, J.S. Communicating Emotion: Linking Affective Prosody and Word Meaning. *J. Exp. Psychol. Hum. Percept. Perform.* **2008**, *34*, 1017.
104. Wilson, D.; Wharton, T. Relevance and Prosody. *J. Pragmat.* **2006**, *38*, 1559-1579.
105. Frühholz, S.; Trost, W.; Kotz, S.A. The Sound of Emotions—Towards a Unifying Neural Network Perspective of Affective Sound Processing. *Neurosci. Biobehav. Rev.* **2016**, *68*, 96-110.
106. Grandjean, D. Brain Networks of Emotional Prosody Processing. *Emot. Rev.* **2021**, *13*, 34-43.
107. Jiang, A.; Yang, J.; Yang, Y. Mmn Responses During Implicit Processing of Changes in Emotional Prosody: An ERP Study Using Chinese Pseudo-Syllables. *Cogn. Neurodyn.* **2014**, *8*, 499-508.
108. Lin, Y.; Fan, X.; Chen, Y.; Zhang, H.; Chen, F.; Zhang, H.; Ding, H.; Zhang, Y. Neurocognitive Dynamics of Prosodic Salience over Semantics During Explicit and Implicit Processing of Basic Emotions in Spoken Words. *Brain Sci.* **2022**, *12*, 1706.
109. Mauchand, M.; Caballero, J.A.; Jiang, X.; Pell, M.D. Immediate Online Use of Prosody Reveals the Ironic Intentions of a Speaker: Neurophysiological Evidence. *Cogn. Affect. Behav. Neurosci.* **2021**, *21*, 74-92.
110. Chen, Y.; Tang, E.; Ding, H.; Zhang, Y. Auditory Pitch Perception in Autism Spectrum Disorder: A Systematic Review and Meta-Analysis. *J. Speech Lang. Hear. Res.* **2022**, *65*, 4866-4886, doi:10.1044/2022_jslhr-22-00254.
111. Zhang, L.; Xia, Z.; Zhao, Y.; Shu, H.; Zhang, Y. Recent Advances in Chinese Developmental Dyslexia. *Annu. Rev. Linguist.* **2023**, *9*, 439-461, doi:10.1146/annurev-linguistics-030421-065648.

112. Zhang, M.; Xu, S.; Chen, Y.; Lin, Y.; Ding, H.; Zhang, Y. Recognition of Affective Prosody in Autism Spectrum Conditions: A Systematic Review and Meta-Analysis. *Autism* **2022**, *26*, 798-813, doi:10.1177/1362361321995725.
113. Seddoh, S.A. How Discrete or Independent Are''Affective Prosody''and''Linguistic Prosody''? *Aphasiology* **2002**, *16*, 683-692.
114. Ben-David Boaz M.; Gal-Rosenblum Sarah; van Lieshout Pascal H. H. M.; Shakuf Vered Age-Related Differences in the Perception of Emotion in Spoken Language: The Relative Roles of Prosody and Semantics. Journal of Speech, Language, and Hearing Research 2019, 62, 1188–1202, doi:10.1044/2018_JSLHR-H-ASCC7-18-0166.