Speech Communication 160 (2024) 103084

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Visual-articulatory cues facilitate children with CIs to better perceive Mandarin tones in sentences

Ping Tang *, Shanpeng Li, Yanan Shen, Qianxi Yu, Yan Feng

School of Foreign Studies, Nanjing University of Science and Technology, Nanjing, Jiangsu, China

ARTICLE INFO

Keywords:
Mandarin tones
Speech perception
Cochlear implants
Language acquisition
Visual-articulatory cues

ABSTRACT

Children with cochlear implants (CIs) face challenges in tonal perception under noise. Nevertheless, our previous research demonstrated that seeing visual-articulatory cues (speakers' facial/head movements) benefited these children to perceive isolated tones better, particularly in noisy environments, with those implanted earlier gaining more benefits. However, tones in daily speech typically occur in sentence contexts where visual cues are largely reduced compared to those in isolated contexts. It was thus unclear if visual benefits on tonal perception still hold in these challenging sentence contexts. Therefore, this study tested 64 children with CIs and 64 age-matched NH children. Target tones in sentence-medial position were presented in audio-only (AO) or audiovisual (AV) conditions, in quiet and noisy environments. Children selected the target tone using a picture-point task. The results showed that, while NH children did not show any perception difference between AO and AV conditions, children with CIs significantly improved their perceptual accuracy from AO to AV conditions. The degree of improvement was negatively correlated with their implantation ages. Therefore, children with CIs were able to use visual-articulatory cues to facilitate their tonal perception even in sentence contexts, and earlier auditory experience might be important in shaping this ability.

1. Introduction

Cochlear implants (CIs) have made oral speech possible for many children with severe hearing impairments (Ching et al., 2018). Yet, the perception of lexical tones remains challenging since CI devices do not effectively transmit pitch information, the primary acoustic carrier of tones (Vandali & van Hoesel, 2012). Nevertheless, daily conversations typically take place in audiovisual environments, where listeners not only hear speakers' speech sounds (auditory cues) but also simultaneously see their articulatory movements (visual cues). It has been found that Mandarin tonal productions are accompanied by specific articulatory movements, particularly in speakers' facial and head movements (De Menezes et al., 2020; Garg et al., 2019, 2023). While visual-articulatory cues of tones are arguably less salient as compared to those of segments (vowels and consonants) and less influential on perception (Hong et al., 2023), our pioneering study on *tones in isolation* has shown that visual information enhanced children with CIs' tonal perception in noise, with children implanted earlier showing greater benefits (left blank for anonymous review). However, tones in daily conversations typically occur in connected speech, i.e., in sentence contexts, where children with CIs face greater challenges in speech perception and tonal recognition in sentences (Gao et al., 2021; Hong et al., 2019; Tao et al., 2014; Zhang et al., 2018), especially in a noisy environment (Huang et al., 2020). At issue, therefore, is whether children with CIs still show a visual facilitation effect on tonal perception in sentences, where visual-articulatory cues of tones become less salient as compared to those in isolation forms (Attina et al., 2010; Burnham et al., 2006, 2022); if yes, whether such effect is contingent on the age of implantation. Addressing these questions will better inform the role of visual cues in speech perception at the suprasegmental level (tones), and the capacity of children with CIs in visual speech perception.

Mandarin Chinese uses four lexical tones in addition to vowels and consonants to differentiate word meanings, acoustically instantiated by pitch inflections, i.e., level (Tone 1 or T1), rising (T2), dipping (T3), and falling (T4). Lexical tones in isolation forms are also characterized by inherent durational differences, e.g., T3 is the longest, followed by T1 and T2, and T4 is the shortest (Fu & Zeng, 2000). It has been shown that NH children acquire lexical tones early (before age 3) with high accuracy in tonal perception (Wang, Schwartz & Jenkins, 2005; (Tang et al., 2019, 2019; Zheng et al., 2009)), while children with CIs face significant

* Corresponding author.
E-mail address: ping.tang@njust.edu.cn (P. Tang).

https://doi.org/10.1016/j.specom.2024.103084
Received 15 December 2023; Received in revised form 9 April 2024; Accepted 12 May 2024
Available online 17 May 2024

challenges in tonal perception especially in noisy environments (see (Chen & Wong, 2017; Tan et al., 2016) for reviews). For example, the tonal perception accuracy of NH children exceeded 90% in both quiet and noisy environments (at a 0 dB Signal-to-Noise Ratio, or SNR; (Mao & Xu, 2017)), while the CI group's accuracy ranged from 70% to 80% in quiet and decreased to around 60% in noise (chance level = 50%; (Chen & Wong, 2017; Mao & Xu, 2017)). Despite these challenges, it has been found that early implantation typically leads to better tonal perception in both quiet and noisy environments, as reflected by the negative correlations observed between children's tonal perception accuracy and their implantation ages (Mao & Xu, 2017).

Speech perception has been well-recognized as a bimodal process in which listeners integrate information from both auditory and visual channels into comprehension, and dynamically allocate more attentional resources to the visual channel when auditory information is less clear (Buchan, 2011; Jesse & Kaplan, 2019; McGurk & MacDonald, 1976). Visual cues, particularly lip movements, substantially enhance the perception of vowels and consonants for children with CIs, especially in noisy environments (Stevenson et al., 2017). Early implantation also boosts this integration ability (Schorr et al., 2005). However, unlike vowels and consonants which exhibit explicit and strong visual-articulatory associations, lexical tone variations hinge on glottal and sub-glottal activities independent of vocal tract configurations (Wang et al., 2020). Nonetheless, the computer-vision analysis still reveals that during tonal productions, the speaker's facial and head movements are aligned with the specific spatial and temporal pitch movement trajectories of different tones (De Menezes et al., 2020; Garg et al., 2019, 2023). Some of these movements are believed to be physiologically motivated, relating to the movements of the laryngeal muscles controlling the vibration of vocal folds (Wang et al., 2020). Specifically, T1 involves minimal movements of the above areas, T2 and T3 end with downward or upward movements of the eyebrows and head, and T4 is associated with a specific lip-closing movement (Garg et al., 2019). Additionally, the duration of tonal articulatory movements also varies across tones, reflecting the inherent duration of tones in isolation mentioned before (Fu & Zeng, 2000).

While visual-articulatory associations of tones are less explicit and less transparent as compared to those of segments (Hong et al., 2023), visual cues still play a complementary role in tonal perception especially when auditory cues are unreliable or unavailable (Wang et al., 2020). For example, visual cues have been found to enhance tonal perception for NH listeners under high-noise conditions (SNR = −12 dB; (Li et al., 2022)) and CI-simulated conditions (Smith & Burnham, 2012). In our pioneering study, we examined the role of visual cues in the perception of Mandarin tones in *isolation forms* by 3–7-year-old children with CIs and age-matched NH controls (left blank for anonymous review). Our results demonstrated that while NH children showed no perceptual difference between audio-only (AO) and audiovisual (AV) conditions, children with CIs improved their tonal recognition accuracy from AO to AV condition, though such improvement was only evident in noise (SNR = 0 dB). Furthermore, the extent of visual benefits (i.e., changes in accuracy from AO to AV condition) was negatively correlated with the implantation age. These results, for the first time, demonstrated that children with CIs can use visual cues to facilitate their tonal perception in auditory-challenging environments, and the ability to integrate visual information into tonal identification is more advanced for children implanted earlier.

In everyday conversations, tones are more commonly produced in sentence contexts, where tonal/word recognition in sentences has been demonstrated to pose greater challenges for children with CIs (Gao et al., 2021; Hong et al., 2019; Tao et al., 2014; Zhang et al., 2018), especially in noisy environments (Huang et al., 2020). For example, it has been found that while children with CIs can correctly recognize sentences (accuracy: 77%) in a quiet environment, their recognition rate in noise (SNR: 5 dB) dropped to 50% (Tao et al., 2014). Despite these challenges, children with CIs' speech perception ability in sentences is correlated with factors such as implantation age, CI experience (Hong et al., 2019), and tonal identification ability in citation form (Hong et al., 2019; Tao et al., 2014). However, so far there has been no study examining whether and how these children were able to incorporate visual information into tonal perception in sentence contexts, where visual articulatory features of tones are less distinct as compared to those in isolation (Attina et al., 2010; Burnham et al., 2006, 2022). For instance, relative to the isolation form, tonal productions in sentence-medial positions demonstrated a reduced magnitude of articulatory movements (Attina et al., 2010; Burnham et al., 2006, 2022), and a diminished, or even absent, durational distinction across tones (Peng, 2006; Wu et al., 2023; Yang et al., 2017). It thus raises questions about whether children with CIs can still utilize the limited visual information to aid their perception of tones in sentences, especially in noisy environments, and if such ability is contingent on the implantation age. Exploring these issues would extend our understanding of the multimodal speech perception theory from segmental level to suprasegmental level, and from isolated speech contexts to connected speech settings. It also informs the role of early auditory experience in the development of multimodal speech perception abilities.

Therefore, this study examined the effect of visual information on the perception of Mandarin tones in sentence contexts by children with CIs in both quiet and noisy environments. We first asked (1) do visual cues enhance children with CIs' tonal perception in sentence contexts. If yes, we further asked (2) does earlier implantation facilitate better use of visual cues in tonal perception? Based on our previous findings on tonal perception in isolation forms, we predicted that: (1) children with CIs would show an improved tonal perception accuracy in audiovisual as compared to audio-only settings, particularly in noisy environments, and (2) children implanted earlier would exhibit greater benefits from visual cues in tonal perception.

2. Methods

2.1. Participants

A total of 64 4–7-year-old prelingually deaf children with CIs (range: 38–85 months; mean: 60.70 months; SD: 13.82 months) and 64 age-match (in month) NH controls (range: 38–85 months; mean: 60.77 months; SD: 13.92 months) were recruited. The implantation age of children with CIs ranged from 9 to 83 months (mean: 33.58 months; SD: 18.05 months). All children were recruited from rehabilitation centers (CI group) and kindergartens (NH group) in Beijing and Hebei province, where standard Mandarin Chinese was used in teaching practices and daily conversation. According to oral reports from teachers in the kindergartens, the NH children did not have any speech or hearing difficulties.

2.2. Materials

Target words included 72 monosyllabic words comprising 36 minimal pairs of tones, e.g., "shu1" *book* and "shu3" *mouse*. The 36 minimal pairs constructed all six types of tonal contrasts: T1T2, T1T3, T1T4, T2T3, T2T4, and T3T4, with each tonal contrast including six minimal pairs. Monosyllabic words instead of disyllabic or multisyllabic words were used as they were able to construct a relatively full set of minimal tonal pairs. All target words were picturable and frequent in four-year-olds' language input according to the Tong Corpus (Deng & Yip, 2018) from the CHILDES database (MacWhinney, 2000). Each target word was yoked with a picture depicting the content of the word.

All target words were produced in random order by a female native speaker within a carrier sentence: "qing3 zhao3 dao4 _ zhe4 zhang1 tu2" *Please find _ this picture*. The speaker was asked to produce all target sentences as naturally as possible at her normal speaking rate. During speech production, the upper part of the speaker (neck and above) was audiovisually recorded using a SONY HXR-NX100 video camera. An

external condenser microphone was connected to the camera to enhance the audio quality. All target sentences were recorded in a random order. The production of each sentence was cut into a single MP4 video clip (resolution 1920 × 1080), with a one-second interval before and after the articulation of each utterance during which the speaker's face was at rest (please see Appendix 2 for a computer vision analysis of the current video stimuli).

Acoustic features including pitch and duration of target words were extracted and analyzed to ensure that target tones carried desired acoustic patterns (Fig. 1). The pitch contours match the classic description of tones in sentence contexts (Yip, 2002). A one-way ANOVA was performed on the tonal duration across categories, and the results showed that there were no significant durational differences across tones: F (3, 76) = 0.173, p = 0.915. This implies that the durational distinction of tones in the current stimuli was reduced, consistent with previous research showing that the durational differences of Mandarin tones are largely reduced in non-final positions (Peng, 2006; Wu et al., 2023; Yang et al., 2017).

2.3. Procedure

All children were tested individually in quiet rooms at rehabilitation centers or kindergartens. The background noise level of these rooms was below 40 dB, measured by a sound level meter before each test. The experiment consisted of two sessions for quiet and noisy environments respectively. Due to the limited attention and concentration span particularly for young children with CIs, each child attended the two sessions on two separate days, with a counterbalanced order across participants (Fig. 2). Each session lasted approximately 10 min. The experiment was programmed, presented, and conducted using *PsychoPy* (Peirce, 2007).

The quiet session began with a familiarization phase, followed by a test phase. During the familiarization phase, each child was presented with all target pictures, one at a time, and asked to name them using monosyllabic words. Most children were able to name most of the pictures using the corresponding target words. On very rare occasions, children used non-target words to name the pictures. For instance, if a child used the disyllabic word "lao3 shu3″ *old mouse* to name the picture for "shu3" *mouse*, the experimenter would correct them and ask him/her to name it again. At the end of the familiarization phase, the experimenter would show any incorrectly named picture again and ask the child to name it one more time. All children were able to successfully produce the target word after one correction and remembered its target name by the end of the familiarization phase.

The test phase consisted of two blocks for AO and AV conditions respectively, with a counterbalanced order across participants. Each block began with a practice trial and six test trials with different tonal contrasts. The procedures for practice and test trials were identical except that participants' responses in practice trials were not included in further analysis. In each trial, a target sentence (e.g., "qing3 zhao3 dao4 shu1 zhe4 zhang1 tu2" *Please find book this picture*; underline indicates the target word) was played either in audio-only forms (for the AO condition) or audiovisual forms (for the AV condition). For example, in the AO condition, only the speech sound of the target sentence was played (via an Edifier HECATE G2000 speaker at approximately 70 dB), with a cartoon image of a loudspeaker being presented on the screen; in the AV condition, the speech sound and the video of the speaker were both presented. Then, two pictures corresponding to the target word (e.g., book) and its tone-contrast counterpart (e.g., mouse) were presented side by side on the computer screen, with the positions of the target and non-target pictures randomized across trials. The child was required to point to the corresponding picture and his/her response was recorded by the experimenter through button pressing. The choice of the target word from each minimal pair was counterbalanced across participants, and each participant was exposed to target words carrying all four lexical tones within each block.

The procedure of the noise session was similar to that of the quiet session, but the audio stimuli in both AO and AV conditions in the noisy session were superimposed with an eight-talker babble noise at a 0 dB SNR. Such noise level/type was used to keep consistent with our previous study on audiovisual perception of tones in citation form (left blank for anonymous review), which was selected based on the results of our pilot experiment and previous evidence that children with CIs' tonal perception performance is around chance level at 0 dB SNR (Mao & Xu, 2016). The entire experiment followed a two-by-two design, including two environments (quiet and noisy) combined with two conditions (AO and AV), which resulted in four condition combinations for each participant: AO in quiet, AV in quiet, AO in noise, and AV in noise. Note that four different sets of words were used across conditions to avoid any word repetition and potential memory effects. The selection of the four word-sets (from the six sets in total) was counterbalanced across participants.

2.4. Analysis

Children's response accuracy (i.e., correct or incorrect) was collected and analyzed. To test our first hypothesis that if children with CIs would show an improved tonal recognition accuracy from AO to AV condition, two generalized linear-mixed effect models (for NH and CI groups in separate) were conducted on the response accuracy across conditions and environments. This was conducted using the *glmer* function of the R package "lme4" (Bates et al., 2015). Statistical significances were identified using likelihood-ratio tests, implemented by the *mixed* function of the R package "afex" (Singmann et al., 2023). When a significant finding was identified, Tukey-HSD post-hoc comparisons were conducted using the *emmeans* function of the R package "emmeans" (Lenth, 2022).

To address our second hypothesis that whether the degree of visual benefits was correlated with children's implantation ages, the difference score of tonal perception accuracy between AO and AV conditions (AV - AO) was calculated for each CI participant. Two Pearson-correlation tests (in quiet and noise separately) were performed between the difference scores and children's implantation ages, implemented by the *cor.test* function of the R package "stats" (R (Core Team, 2022))[1]

3. Results

3.1. Tonal recognition accuracy

Fig. 3 illustrates NH and CI groups' tonal recognition accuracy in AO and AV conditions across environments. Two separate generalized linear mixed-effects models were performed on the recognition accuracy for NH and CI groups, with two fixed factors "Condition" (AO and AV) and "Environment" (Quiet and Noise), and two random factors "Participant" and "Target word" in each model.[2]

The model for the NH group showed a significant effect for "Environment" ($\chi^2(1) = 5.09, p = 0.024$) but not for "Condition" ($\chi^2(1) = 1.41$, p = 0.235) or the interaction of "Environment × Condition" ($\chi^2(1)$ = 0.99, p = 0.320). These results implied that there was no evidence supporting that NH children's tonal perception accuracy differed between AO and AV conditions.

The model for the CI group reported a significant effect of

[1] As an anonymous reviewer has pointed out, there might be other contribution factors that could influence the degree of visual benefits. Therefore, based on the demographic information we have in Appendix 1, we also performed an exploratory regression analysis on visual benefits, with predictors of Implantation age, CI experience and Implantation type (Unilateral CI, Bilateral CI and Bimodal). Please see Appendix 2 for detailed results.

[2] R code for each model: Accuracy ~ Condition * Environment + (1|Participant) + (1|Target word), family = binomial (link = "logit").

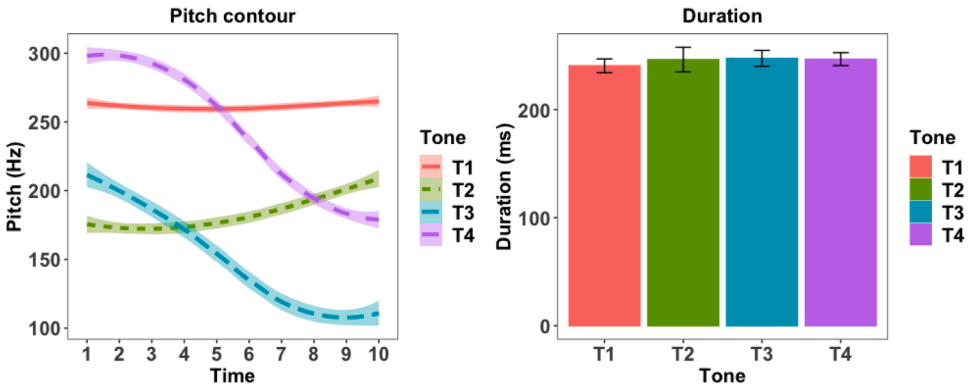

Fig. 1. Pitch contours and duration of lexical tones from target words.

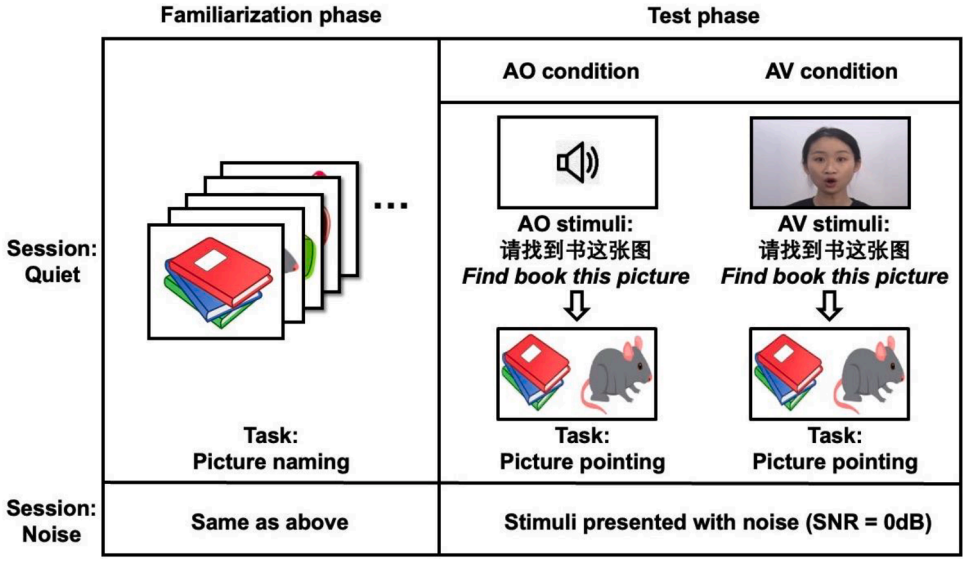

Fig. 2. A paradigm of the experimental procedure.

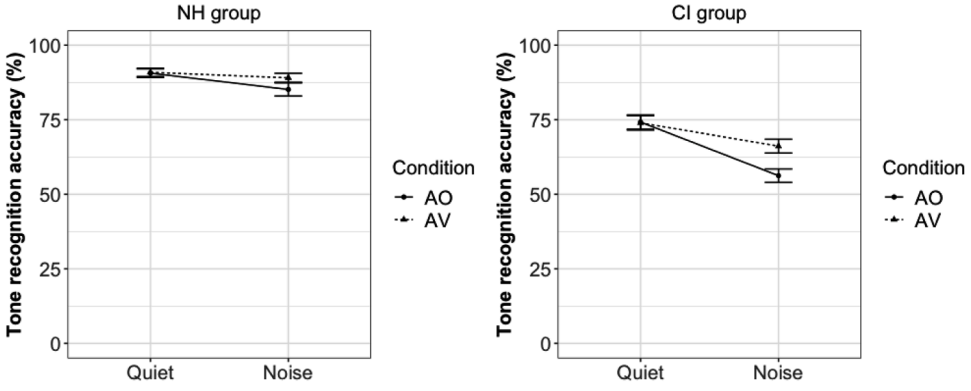

Fig. 3. NH and CI groups' tonal recognition accuracy in AO and AV conditions.

"Environment" ($\chi^2(1) = 29.55$, $p < 0.001$) and a significant interaction of "Environment × Condition" ($\chi^2(1) = 3.93$, $p = 0.047$), while the main effect of "Condition" was not significant ($\chi^2(1) = 3.46$, $p = 0.063$). These results suggest that the CI group's perception accuracy differed between the AO and AV conditions, while such differences might be contingent on the environment. A Tukey-HSD post-hoc test performed on the interaction further revealed no significant accuracy difference between AO and AV conditions in the quiet environment (AO – AV: $\beta = 0.003$, SE $= 0.032$, $z = 0.084$, $p = 0.933$). However, in the noisy environment, the AV condition exhibited higher accuracy than the AO condition (AO – AV: $\beta = -0.102$, SE $= 0.036$, $z = -2.879$, $p = 0.004$).

3.2. Correlation between implantation age and visual benefit

Fig. 4 illustrates the relationships between implantation ages and visual benefits (difference scores) in quiet and noisy environments.

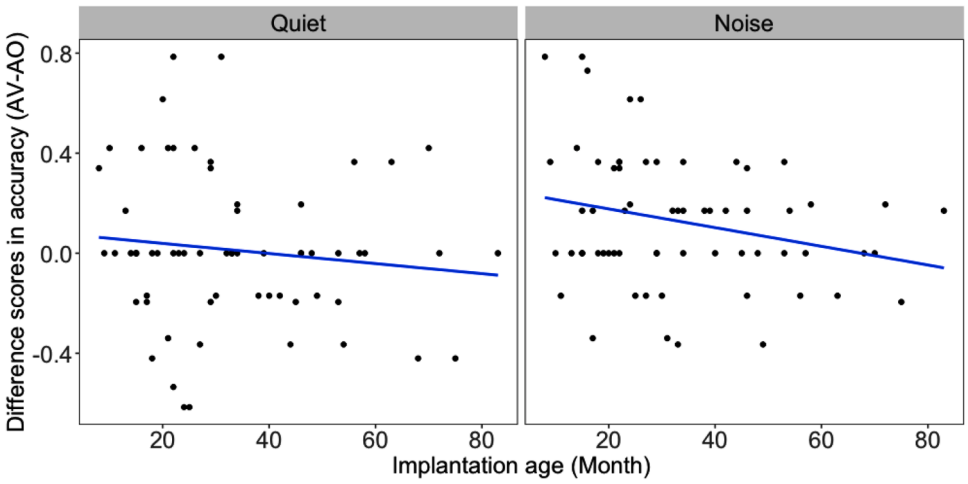

Fig. 4. Correlations between implantation ages and visual benefits (difference scores in accuracy between conditions: AV-AO).

Pearson correlation tests revealed a negative correlation between the two factors in the noisy environment ($r = -0.26$, t (62) $= -2.090$, $p = 0.041$), but no significant correlation was found in the quiet environment ($r = -0.12$, t (62) $= -0.952$, $p = 0.345$).

4. Discussion

This study explored the effect of visual-articulatory cues on the perception of Mandarin tones in sentence contexts by children with CIs. The results showed that visual cues significantly improved children with CIs' tonal recognition accuracy in sentences, although such benefits only existed in a noisy environment. Furthermore, the implantation age was negatively correlated with the degree of visual benefits in noisy environments, suggesting that children implanted earlier benefited more from visual cues in tonal perception in sentences.

The current results aligned with our previous research of visual benefits for tonal perception in isolation forms and extended these findings to tones in connected speech (left blank for anonymous review). The combined results thus revealed that in auditory-challenging situations, visual cues might serve as a robust and reliable source for children with CIs to perceive tones better across different contexts. Furthermore, the consistently observed benefits of early implantation highlight its important role in the development of multimodal perception abilities of tones across various contexts.

However, it should be noted that in the current results, the NH group did not demonstrate any significant visual benefit on tonal perception in either quiet or noisy environments, and the CI group did not exhibit such a benefit in the quiet environment neither. This observation might be attributed to the fact that, unlike segments, the visual information of tones serves only as a secondary cue for perception, playing a complementary role to tonal perception when auditory information is insufficient or unclear, such as in a noisy environment for the CI group (Hong et al., 2023). In situations where the auditory information is clear and abundant, as in the current quiet and noise settings for the NH group, visual cues become largely redundant and did not provide additional benefits to tonal perception (Wang et al., 2020). This also calls for future studies to investigate how NH children might integrate visual cues into tonal perception in challenging scenarios and how this integration ability develops with age.

Our results also revealed an early implantation advantage in better using visual cues to improve tonal perception abilities in sentence contexts. This advantage could be attributed to two main factors. On the one hand, early access to speech sounds enables children with CIs to better detect visual-auditory correspondences and associate visual cues to phonetic information (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999; Rosenblum et al., 1997). This might further facilitate the development of a multi-sensory (audiovisual) representation of lexical tones, helping children implanted earlier to better access the visual representation of tones during perception and thus gain more benefits from visual cues (Hasson et al., 2007; Smith et al., 2013). On the other hand, the process of audiovisual perception entails a dynamic allocation of cognitive/attentional resources to audio and visual channels in response to environmental factors such as the availability of auditory information (Buchan, 2011; Jesse & Kaplan, 2019). Early auditory deprivation has been demonstrated to be associated with delayed development of various cognitive abilities (see (Lieu et al., 2020), for a review), while early implantation contributes to better cognitive development in general (see (Marschark et al., 2019) for a review). This suggests that children implanted earlier may possess a stronger ability to actively shift attentional resources to visual information under degraded auditory conditions, thereby enhancing their tonal perception.

The current study therefore contributes to the multimodal speech perception theory for children with CIs in several aspects. First, it extends the research scope from the well-studied segmental level to the much-understudied suprasegmental level. We demonstrated that visual cues, despite traditionally being considered to be less salient and less influential for tonal perception as compared to segments, still play a supporting role in comprehending these acoustically nuanced pitch-based tonal variations, even when they are embedded within connected speech where visual cues are further reduced. Therefore, the role of visual information in speech comprehension might be more profound than traditionally thought. Second, our results also highlight the remarkable capacity of children with CIs in visual speech perception. Such ability might be related to the cross-modal plasticity after auditory deprivation, which enhances these children's visual motion and change detection abilities as well as connections between auditory and visual cortexes (see (Kral & Sharma, 2023) for a review), thus enabling them to use visual cues more effectively to compensate for auditory limitations. Thirdly, our results also underpin the important role of early auditory experience in the development of visual speech perception abilities. This highlights the importance of newborn hearing screening and timely intervention strategies, which have also been demonstrated to be critical for the tonal development of children with CIs in general (Tang et al., 2019, 2021).

Two additional key questions remain. First, do visual cues provide an overall benefit to pitch perception in general (Esteve-Gibert & Guellaï, 2018), or do they specifically enhance lexical tone perception as a unique multisensory phonological category? This calls for future studies to examine the effect of visual cues in the perception of other pitch-related phenomena such as intonation, as well as the interaction

between tone and intonation, to disentangle the visual benefits on tonal and pitch perception. Second, can specific training improve children with CIs' visual perception ability for tones? This also calls for future studies to identify potential training methods and examine their impact on these children's visual perception of tones in diverse speech contexts (Chen & Massaro, 2008).

There are also several limitations of this study. First, the current stimuli only used target words embedded in a fixed sentence context, and it was unclear whether these findings were generalizable to other contexts or more naturalistic, daily speech conversation contexts. This calls for future studies to further examine children with CIs' performance in more natural sentence contexts to better explore the role of visual cues on their tonal perception in their daily speech settings. Second, it has been found that, apart from implantation age, many other factors might also influence the tonal perception performance by children with CIs, such as phonological awareness (Zhang et al., 2022), auditory working memory (Tao et al., 2014), musical training experience (Cheng et al., 2018), Speech training experience (Zhang et al., 2023), implantation type (Cheng et al., 2018; Yuen et al., 2009), etc. Furthermore, socioeconomic factors such as maternal education, family income, and accessibility to speech therapy have also been demonstrated to play critical roles in general language development by children with CIs (Szagun & Stumper, 2012; Sharma et al., 2017). Therefore, future studies could also examine the contribution of the above factors in children with CIs' ability to integrate audiovisual cues into tonal perception, to draw a whole picture of the tonal development of this population.

In conclusion, our study reveals that children with CIs were able to use visual-articulatory cues to improve their tonal perception in connected speech and noisy environments. Thus, visual cues not only impact on the perception of segments but also contribute to the perception of lexical tones, and early language experience might play a critical role in shaping children's visual speech perception abilities.

CRediT authorship contribution statement

Ping Tang: Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Shanpeng Li:** Methodology, Investigation, Conceptualization. **Yanan Shen:** Visualization, Methodology, Investigation, Formal analysis. **Qianxi Yu:** Methodology, Formal analysis. **Yan Feng:** Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

The research is funded by The National Social Science Fund of China (20CYY012).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.specom.2024.103084.

References

Attina, V., Gibert, G., Vatikiotis-Bateson, E., Burnham, D., 2010. Production of Mandarin lexical tones: auditory and visual components. Auditory-Visual Speech Processing, 2010.
Bates, D., Mächler, M., Bolker, B., Walker, S., Christensen, R.H., Singmann, H., Dai, B., 2015. lme4: Linear mixed-Effects Models Using Eigen and S4, 1–7. R package version 1, p. 2014.
Buchan, J.N., 2011. *Cognitive Resources in Audiovisual Speech Perception* (Doctoral Dissertation. Queen's University.
Burnham, D., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H., Ciocca, V., Morris, R., Haszard, Hill, H., Vignali, G., Bollwerk, S., Tam, H., Jones, C., 2006. The perception and production of phones and tones: the role of rigid and non-rigid face and head motion. In: Yehia, H (Ed.), Proceedings of the 7th International Seminar on Speech Production. CEFALA, Brazil, pp. 1–8.
Burnham, D., Vatikiotis-Bateson, E., Vilela Barbosa, A., Menezes, J.V., Yehia, H.C., Morris, R.H., Vignali, G., Reynolds, J., 2022. Seeing lexical tone: Head and face motion in production and perception of Cantonese lexical tones. Speech. Commun. 141, 40–55.
Chen, T.H., Massaro, D.W., 2008. Seeing pitch: visual information for lexical tones of Mandarin-Chinese. J. Acoust. Soc. Am. 123 (4), 2356–2366.
Chen, Y., Wong, L.L., 2017. Speech perception in Mandarin-speaking children with cochlear implants: a systematic review. Int. J. Audiol. 56 (sup2), S7–S16.
Cheng, X., Liu, Y., Shu, Y., Tao, D.-D., Wang, B., Yuan, Y., Galvin, J.J., Fu, Q.-J., Chen, B., 2018. Music training can improve music and speech perception in pediatric Mandarin-speaking cochlear implant users. Trends. Hear. 22, 1–12.
Cheng, X., Liu, Y., Wang, B., Yuan, Y., Galvin, J.J., Fu, Q.-J., Shu, Y., Chen, B., 2018. The benefits of residual hair cell function for speech and music perception in pediatric bimodal cochlear implant listeners. Neural Plast. 2018, 1–11.
Ching, T.Y., Dillon, H., Leigh, G., Cupples, L., 2018. Learning from the Longitudinal Outcomes of Children with Hearing Impairment (LOCHI) study: summary of 5-year findings and implications. Int. J. Audiol. 57 (sup2), S105–S111.
De Menezes, J.V.P., Cantoni, M.M., Burnham, D., Barbosa, A.V., 2020. A method for lexical tone classification in audio-visual speech. J. Speech Sci.
Deng, X., Yip, V., 2018. A multimedia corpus of child Mandarin: the Tong corpus. J. Chin. Linguistics 46 (1), 69–92.
Esteve-Gibert, N., Guellaï, B., 2018. Prosody in the auditory and visual domains: a developmental perspective. Front. Psychol. 9, 338.
Fu, Q.J., Zeng, F.G., 2000. Identification of temporal envelope cues in Chinese tone recognition. Asia Pacific J. Speech Lang. Hear. 5 (1), 45–57.
Gao, Q., Wong, L.L.N., Chen, F., 2021. A review of speech perception of Mandarin-speaking children with cochlear implantation. Front. Neurosci. 15, 773694.
Garg, S., Hamarneh, G., Jongman, A., Sereno, J.A., Wang, Y., 2019. Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories. Speech. Commun. 113, 47–62.
Garg, S., Hamarneh, G., Sereno, J., Jongman, A., Wang, Y., 2023. Different facial cues for different speech styles in Mandarin tone articulation. Front. Commun. (Lausanne) 8, 1148240.
Hasson, U., Skipper, J.I., Nusbaum, H.C., Small, S.L., 2007. Abstract coding of audiovisual speech: beyond sensory representation. Neuron 56 (6), 1116–1126.
Hong, S., Wang, R., Zeng, B., 2023. Incongruent visual cues affect the perception of Mandarin vowel but not tone. Front. Psychol. 13, 971979.
Hong, T., Wang, J., Zhang, L., Zhang, Y., Shu, H., Li, P., 2019. Age-sensitive associations of segmental and suprasegmental perception with sentence-level language skills in Mandarin-speaking children with cochlear implants. Res. Dev. Disabil. 93, 103453.
Huang, W., Wong, L.L.N., Chen, F., Liu, H., Liang, W., 2020. Effects of fundamental frequency contours on sentence recognition in Mandarin-speaking children with cochlear implants. J. Speech Lang. Hear. Res. 63 (11), 3855–3864.
Jesse, A., Kaplan, E., 2019. Attentional resources contribute to the perceptual learning of talker idiosyncrasies in audiovisual speech. Attention Percept. Psychophys. 81 (4), 1006–1019.
Kral, A., Sharma, A., 2023. Crossmodal plasticity in hearing loss. Trends Neurosci.
Kuhl, P.K., Meltzoff, A.N., 1982. The bimodal perception of speech in infancy. Science (1979) 218 (4577), 1138–1141.
Lenth, R. (2022). emmeans: estimated marginal means, aka least-squares means. R package version 1.7.4-1, <https://CRAN.R-project.org/package=emmeans>.
Li, M., Chen, X., Zhu, J., Chen, F., 2022. Audiovisual mandarin lexical tone perception in quiet and noisy contexts: the influence of visual cues and speech rate. J. Speech. Lang. Hear. Res. 65 (11), 4385–4403.
Lieu, J.E., Kenna, M., Anne, S., Davidson, L., 2020. Hearing loss in children: a review. JAMa 324 (21), 2195–2205.
Mao, Y., Xu, L., 2017. Lexical tone recognition in noise in normal-hearing children and prelingually deafened children with cochlear implants. Int. J. Audiol. 56 (sup2), S23–S30.
Marschark, M., Duchesne, L., Pisoni, D., 2019. Effects of age at cochlear implantation on learning and cognition: a critical assessment. Am. J. Speech. Lang. Pathol. 28 (3), 1318–1334.
McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264 (5588), 746–748.
Patterson, M.L., Werker, J.F., 1999. Matching phonetic information in lips and voice is robust in 4.5-month-old infants. Infant Behav. Dev. 22 (2), 237–247.
Peirce, J.W., 2007. PsychoPy—psychophysics software in Python. J. Neurosci. Methods 162 (1–2), 8–13.
Peng, G., 2006. Temporal and tonal aspects of Chinese syllables: a corpus-based comparative study of Mandarin and Cantonese. J. Chin. Linguistics 34 (1), 134.

Core Team, R., 2022. R: A language and Environment For Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL. https://www.R-project.org/.
Rosenblum, L.D., Schmuckler, M.A., Johnson, J.A., 1997. The McGurk effect in infants. Percept. Psychophys. 59 (3), 347–357.
Schorr, E.A., Fox, N.A., van Wassenhove, V., Knudsen, E.I., 2005. Auditory-visual fusion in speech perception in children with cochlear implants. Proc. Natl. Acad. Sci. 102 (51), 18748–18750.
Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M., 2023. afex: Analysis of Factorial Experiments. R package version 1.3-0.<https://CRAN.R-project.org/package=afex>.
Smith, E., Duede, S., Hanrahan, S., Davis, T., House, P., Greger, B., 2013. Seeing is believing: neural representations of visual stimuli in human auditory cortex correlate with illusory auditory perceptions. PLoS. One 8 (9), e73148.
Stevenson, R., Sheffield, S.W., Butera, I.M., Gifford, R.H., Wallace, M., 2017. Multisensory integration in cochlear implant recipients. Ear Hear. 38 (5), 521.
Tan, J., Dowell, R., Vogel, A., 2016. Mandarin lexical tone acquisition in cochlear implant users with prelingual deafness: a review. Am. J. Audiol. 25 (3), 246–256.
Tang, P., Yuen, I., Xu Rattanasone, N., Gao, L., Demuth, K., 2019. Acquisition of weak syllables in tonal languages: acoustic evidence from neutral tone in Mandarin Chinese. J. Child Lang. 46 (1), 24–50.
Tang, P., Yuen, I., Xu Rattanasone, N., Gao, L., Demuth, K., 2019. The acquisition of phonological alternations: the case of the Mandarin tone sandhi process. Appl. Psycholinguist. 40 (6), 1495–1526.
Tang, P., Yuen, I., Xu Rattanasone, N., Gao, L., Demuth, K., 2019. The acquisition of Mandarin tonal processes by children with cochlear implants. J. Speech Lang. Hear. Res. 62 (5), 1309–1325.
Tang, P., Yuen, I., Xu Rattanasone, N., Gao, L., Demuth, K., 2021. Longer cochlear implant experience leads to better production of Mandarin tones for early implanted children. Ear Hear. 42 (5), 1405–1411.
Tao, D., Deng, R., Jiang, Y., Galvin, J.J., Fu, Q.-J., Chen, B., 2014. Contribution of auditory working memory to speech understanding in Mandarin-speaking cochlear implant users. PLoS. One 9 (6), e99096.
Vandali, A.E., van Hoesel, R.J., 2012. Enhancement of temporal cues to pitch in cochlear implants: effects on pitch ranking. J. Acoust. Soc. Am. 132 (1), 392–402.
Wang, Y., Sereno, J.A., Jongman, A., 2020. Multi-modal perception of tone. Speech Percept. Prod. Acquisit. 159–173.
Wu, Y., Adda-Decker, M., Lamel, L., 2023. Mandarin lexical tone duration: impact of speech style, word length, syllable position and prosodic position. Speech. Commun. 146, 45–52.
Yang, J., Zhang, Y., Li, A., Xu, L., 2017. On the duration of mandarin tones. INTERSPEECH, pp. 1407–1411.
Yip, M., 2002. Tone. Cambridge University Press.
Yuen, K.C.P., Cao, K.-L., Wei, C.-G., Luan, L., Li, H., Zhang, Z.-Y., 2009. Lexical tone and word recognition in noise of Mandarin-speaking children who use cochlear implants and hearing aids in opposite ears. Cochlear. Implants Int. 10 (1), 120–129. Suppl.
Zhang, H., Ma, W., Ding, H., Peng, G., Zhang, Y., 2022. Phonological awareness and working memory in Mandarin-speaking preschool-aged children with cochlear implants. J. Speech Lang. Hear. Res. 65 (11), 4485–4497.
Zhang, H., Ma, W., Ding, H., Zhang, Y., 2023. Sustainable benefits of high variability phonetic training in mandarin-speaking kindergarteners with cochlear implants: evidence from categorical perception of lexical tones. Ear Hear. 44 (5), 990–1006.
Zhang, L., Wang, J., Hong, T., Li, Y., Zhang, Y., Shu, H., 2018. Mandarin-speaking, kindergarten-aged children with cochlear implants benefit from natural F0 patterns in the use of semantic context during speech recognition. J. Speech Lang. Hear. Res. 61 (8), 2146–2152.
Zheng, Y., Meng, Z.L., Wang, K., Tao, Y., Xu, K., Soli, S.D., 2009. Development of the Mandarin early speech perception test: Children with normal hearing and the effects of dialect exposure. Ear Hear. 30 (5), 600–612.