Speech Communication 169 (2025) 103199

Contents lists available at ScienceDirect

Speech Communication

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

Effects of harmonicity on Mandarin speech perception in cochlear implant users

Mingyue Shi [a], Qinglin Meng [b,d], Huali Zhou [c,d], Jiawen Li [a], Yefei Mo [e], Nengheng Zheng [a,d,*]

[a] Guangdong Provincial Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, 518052, China
[b] Acoustics Laboratory, School of Physics and Optoelectronics, South China University of Technology, Guangzhou, 510641, China
[c] School of Electronics and Information Engineering, Heyuan Polytechnic, Heyuan, 517000, China
[d] Shenzhen Longgang E.N.T Hospital-South China University of Technology Joint Laboratory for Digital Hearing Healthcare, Shenzhen, 518172, China
[e] Department of Otolaryngology-Head and Neck Surgery, Beijing Friendship Hospital, Capital Medical University, Beijing, 100050, China

ARTICLE INFO

Keywords:
Harmonicity
Speech perception
Mandarin
Cochlear implant

ABSTRACT

Previous research has demonstrated the negligible impact of harmonicity on English speech perception for normal hearing (NH) listeners in quiet environments. This study aims to bridge the gap in understanding the role of harmonicity in Mandarin speech perception for cochlear implant (CI) users. Speech perception in quiet was tested in both CI simulation group and actual CI user group using harmonic and inharmonic Mandarin speech. Furthermore, speech-on-speech perception was tested in NH, CI simulation, and actual CI user groups. For speech perception in quiet, results show that, compared to harmonic speech, inharmonic speech decreased the mean recognition rate for both actual CI user and CI simulation groups by about 10 percentage points. For speech-on-speech perception, all groups (i.e., NH, CI simulation, and actual CI user) performed worse with inharmonic stimuli compared to harmonic stimuli. The findings of this study, along with previous studies in NH listeners, indicate that harmonicity aids target speech recognition for NH listeners in speech-on-speech conditions but not speech perception in quiet. In contrast, harmonicity plays an important role in CI users' Mandarin speech recognition in both quiet and speech-on-speech conditions. However, under speech-on-speech conditions, CI users could only understand target speech at positive SNRs (often > 5 dB), suggesting that their performance depends on the intelligibility of the target speech. The contribution of harmonicity to masking release in CI users remains unclear.

1. Introduction

Speech perception is based on the considerable redundancy present in the speech signal and its physiological and neural representation within the auditory system (Plack, 2018). Generally, natural speech includes various acoustic cues, such as envelopes and fine structures from both temporal and spectral domains, which influence intelligibility in quiet and noisy environments (Oxenham, 2008; Qin and Oxenham, 2003; Goldsworthy, 2019). In the spectral domain, the envelope containing the formants is primarily determined by the resonant characteristics of the vocal tract and plays a crucial role in speech intelligibility. Meanwhile, the fine structure consists of harmonic components at frequencies that are integer multiples of the fundamental frequency (F0). These components or combinations of several components appear as periodic patterns on the basilar membrane when viewed in the temporal domain and provide essential cues for pitch perception (McPherson, 2022). Harmonicity refers to the degree to which a signal is composed of harmonic overtones, while periodicity refers to the degree of similarity of a signal with a temporally shifted version of itself. As strictly periodic signals have a harmonic (discrete) spectrum, harmonicity and periodicity are tightly related concepts. Our study examines the impact of harmonicity on Mandarin speech perception in CI users under quiet and noisy conditions.

The effects of harmonic-related spectral fine structure on auditory perception in quiet conditions have been tested in listeners with normal hearing (NH) using various tasks, including pitch-related tasks (McPherson and McDermott, 2018) and speech intelligibility tasks (Popham et al., 2018; Lentz et al., 2022), employing harmonic, inharmonic, and whispered stimuli. For instance, McPherson and McDermott (2018) found that for tasks involving the direction of pitch changes (e.g., discriminating musical melodies and speech prosody), listeners tracked the frequency contours of sinusoid components of the stimuli, irrespective of whether the frequencies were harmonic or inharmonic.

* Corresponding author.
E-mail address: nhzheng@szu.edu.cn (N. Zheng).

https://doi.org/10.1016/j.specom.2025.103199
Received 26 June 2024; Received in revised form 25 November 2024; Accepted 17 January 2025
Available online 24 January 2025

In contrast, tasks that required judgments of pitch intervals or voice identity showed substantially impaired performance for inharmonic stimuli. Popham et al. (2018) found that harmonicity is not effective on the intelligibility of single English words and sentences.

However, for noisy conditions, previous studies have demonstrated that harmonicity is beneficial for (1) speech intelligibility of English words in babble noise (Popham et al., 2018), (2) detection and discrimination of a pitch signal in stationary noise (McPherson et al., 2022; Rajappa et al., 2023), and (3) segregating individual sources from a mixture in NH listeners (Popham et al., 2018; McPherson et al., 2022; Micheyl and Oxenham, 2010; de Cheveigné et al., 1997). Specifically, Popham et al. (2018) found harmonicity is beneficial for speech intelligibility of English words in babble noise but not in speech-shaped noise (SSN). McPherson et al. (2022) and Rajappa et al. (2023) demonstrated that harmonicity improves the detection of synthesized complex-tones and natural speech syllables, as well as the discrimination of pitch, melody, English vowels, and Mandarin tones in stationary noise. Stream segregation, which involves separating and recognizing overlapping sounds, presents a more complex challenge. Popham et al. (2018) demonstrated that harmonicity aids in the grouping and streaming of speech. Additionally, they found that whispered manipulation (i.e., replacing harmonic excitation with noise) significantly impaired the ability to segregate and stream concurrent sentences, highlighting additional benefits of the discrete frequencies produced by harmonic excitation. Recently, a speech-on-speech experiment in McPherson et al. (2022) found that participants more accurately recognized target speech under the harmonic-on-harmonic condition than under the inharmonic-on-inharmonic condition.

Less is known about the effects of harmonicity on perception with cochlear implants (CIs). CIs generally enable good speech communication in quiet environments for most users. In clinically-common coding strategies, such as continuous interleaved sampling (CIS) and advanced combination encoder (ACE), the input sound is first filtered into multiple frequency bands (typically 12–24), and then the temporal envelopes are extracted and represented at individual electrodes. The coding strategies currently employed in CIs provide little or no representation of individual harmonics (Oxenham, 2008). Cooper and Roberts (2007) investigated the influence of electrode separation on auditory stream segregation in actual CI users. Although the study did not explicitly focus on harmonicity, the findings suggest that electrode separation affects CI users' frequency resolution, which is fundamental to the segregation of competing sound sources.

Some studies have used a vocoder to simulate auditory perceptions of CI users in NH listeners and indirectly provided some observations of the role of harmonicity in CI perception. For example, Shannon et al. (1995) found that temporal envelopes from four channels were sufficient to convey semantic information in CI simulation experiments with clean speech. Qin and Oxenham (2005) with NH listeners found that temporal envelope cues may not be sufficient for CI users to discriminate changes in F0 and identify simultaneous vowels based on differences in F0. Spitzer et al. (2009) with NH listeners showed that F0 is crucial for CI users in differentiating strong and weak syllables, aiding lexical segmentation. Stickney et al. (2007) found that F0 differences between competing sentences did not improve performance in real and simulated CI processing. This work implies that harmonicity could not be used to facilitate the F0-related masking release for CI users. Steinmetzger and Rosen (2018) showed that the perception of masked speech with simulated and real CIs significantly benefited from masker periodicity.

To the best of our knowledge, the impact of harmonicity on speech recognition with CIs has not been explicitly (i.e., directly) examined using a harmonic–inharmonic comparison paradigm. In this study, we explore the effects of harmonicity on CI users' Mandarin sentence recognition in both quiet and competing speech conditions. Mandarin has four main tones, i.e., high and level, rising, fall-rising, and falling. For example, the speech of words "妈 (mā, mother)", "麻 (má, hemp)", "马 (mǎ, horse)", and "骂 (mà, scold)" is differentiated from each other solely based on the pitch variations within their tones, which lead to different semantic meanings. In contrast, pitch variations in non-tonal languages like English do not alter word meanings. The speech synthesizer developed by Popham et al. (2018) which could disrupt the frequency relationship between harmonics while maintaining the spectrotemporal envelope of original speech was used to generate stimuli for our experiments. In the quiet condition, we measured the intelligibility of harmonic and inharmonic speech sentences. In the competing speech conditions, we mixed two sentences (i.e., speech-on-speech) to compare a harmonic-on-harmonic condition and an inharmonic-on-inharmonic condition. Experiments were carried out in actual and simulated Chinese CI users.

2. Synthesis of inharmonic speech and acoustic analysis

This section presents the construction method of the inharmonic speech used in this study. Furthermore, acoustic analyses were conducted to compare the natural harmonic speech and the constructed inharmonic speech. First, an autocorrelation-based measure (Steinmetzger and Rosen, 2023) was used to quantify the difference in the degree of periodicity between harmonic and inharmonic speech. Then, envelope modulations were compared to verify that the inharmonic speech did not differ from their harmonic counterparts with respect to their modulation spectra.

2.1. Inharmonic speech construction

The voiced segments of natural speech exhibit harmonic structures in their broadband spectrograms, in which the frequency components are integer multiples of the fundamental frequency (F0). Natural speech with harmonic structures in a time frame can be modeled as follows:

$$s[n] = \sum_{k=1}^{N[n]} A_k[n] \cos[2\pi f_k[n] \frac{n}{f_s} + \varphi_k] \tag{1}$$

where $f_k[n] = k f_0[n]$ is the time-varying frequency of the kth harmonic component, n is time index, and $f_0[n]$ is the fundamental frequency. $N[n]$ denotes the number of harmonic components in the current frame, which can be determined based on the signal's frequency range and energy. $A_k[n]$ represents the instantaneous amplitude of the kth harmonic component, f_s is the sampling frequency, and φ_k is the initial phase of the kth harmonic component.

The harmonic structure in natural speech is disrupted to construct inharmonic speech. In this work, this was implemented using an extension (Popham et al., 2018) of the STRAIGHT algorithm (Kawahara and Morise, 2011). The open source code can be accessed at http://mcdermottlab.mit.edu/downloads.html. The harmonic structure was manipulated by jittering the frequencies of the harmonic components while maintaining the spectrotemporal envelope (Popham et al., 2018) as follows.

$$f_k[n] = k f_0[n] + \alpha_k f_0[n] \tag{2}$$

in which $\alpha_k f_0[n]$ denotes the offset applied to the frequency of the kth harmonic component. In this study, α_k was a random variable uniformly distributed over the interval $[-\alpha, +\alpha]$, where α is an adjustable parameter ranging from 0 to 1, used to control the amount of jittering. When $\alpha = 0$, the harmonic structure is maintained. Conversely, when $0 < \alpha \leq 1$, it introduces jittering to the harmonic structure, resulting in inharmonic speech.

Popham et al. (2018) found that the speech naturalness scores and intelligibility plateaued at $\alpha = 0.3$ within the range of $\alpha = 0$ to 0.5, i.e., increasing α beyond 0.3 (e.g., 0.4 or 0.5) did not further impact intelligibility. Based on this finding, we selected $\alpha = 0.3$ as the maximum level of jitter in this study. Sample spectrograms of natural harmonic speech with $\alpha = 0$ and jittered inharmonic speech with $\alpha = 0.3$ are shown in Fig. 1. The sentence was "上个月很多人得了流感 (Many people got the flu last month)". It can be observed that, in contrast to the harmonic speech in Fig. 1(a), the harmonic structure was disrupted in inharmonic speech in Fig. 1(b).

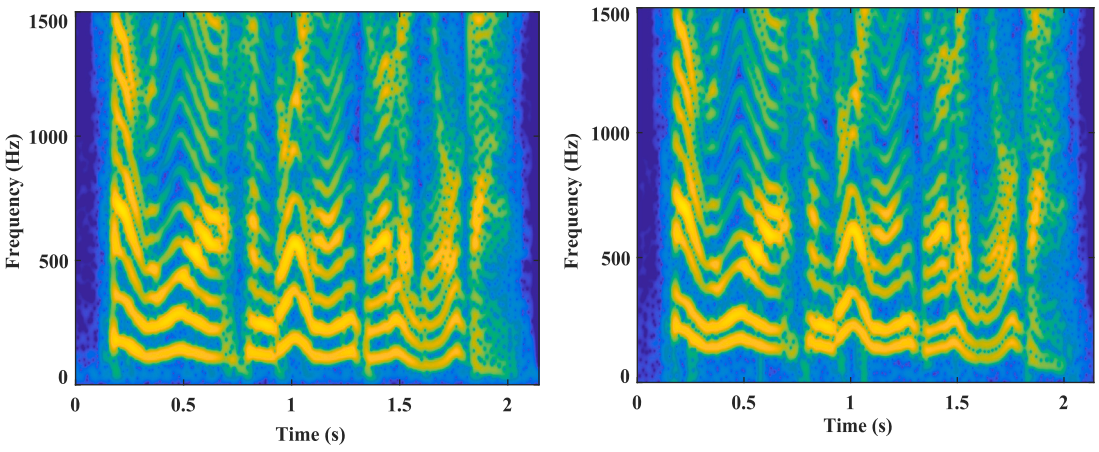


Fig. 1. Sample spectrograms of harmonic (a) and inharmonic (b) speech with $\alpha = 0.3$. The harmonic structure is disrupted in inharmonic speech (b).

2.2. Acoustic analysis

2.2.1. Periodicity level

Disrupting the frequency relationship between harmonics in the frequency domain will lead to changes in the temporal periodicity. To quantify the impact on the temporal periodicity caused by spectral jittering, a periodicity level was calculated based on summary autocorrelation functions (SACFs) (Meddis and Hewitt, 1991; Meddis and O'Mard, 1997). The calculation method is detailed as follows.

Each stimulus was first framed with a 0.05 s Hanning window and a 0.01 s window shift. Then each frame was filtered using a filter bank of 22 Gammatone filters with equivalent rectangular bandwidths ranging from 50 to 8000 Hz (Steinmetzger and Rosen, 2023). The outputs of the filter bank were low-pass filtered, and then the autocorrelation function (ACF) was computed for each channel at various time delays (i.e., time lags). Sample ACFs across the 22 channels for a frame of harmonic and inharmonic speech are given in Fig. 2(a) and (b). For harmonic speech, as shown in Fig. 2(a), the ACFs consistently exhibit a peak at the same time lag across the all 22 channels, and this time lag corresponds to the periodicity of the harmonic speech signal. Conversely, for inharmonic speech, the ACF do not show distinct peaks, as shown in Fig. 2(b).

Autocorrelation functions of 22 channels were averaged into the SACFs, as shown in Fig. 2(c) and (d). For harmonic speech, the SACF (Fig. 2(c)) demonstrates peaks at the same ACF-peak times lags in Fig. 2(a). The first peak corresponds to the F0, and the second peak indicates the second-order harmonic, i.e., 2F0. Conversely, the SACF of inharmonic speech has no regular peak as shown in Fig. 2(d).

Furthermore, the above procedure was applied across the duration of each stimulus, resulting in the spectrographic representations in Fig. 2(e) and (f), with time lags transformed into frequencies. The first peak in SACF spectrograms indicated the F0 contour. As shown in Fig. 2(e), harmonic speech shows a clear F0 contour, which is noticeably more pronounced than that of the inharmonic speech in Fig. 2(f).

Finally, the periodicity level was quantified by computing the average height of the first peak relative to the adjacent trough across each consecutive SACF frame. Specifically, the adjacent trough was determined by finding the peak of the inverted signal (i.e., -SACF) that is closest to the SACF peak. Theoretically, the periodicity level may vary between 0 and 1, the higher the value, the more periodic the temporal waveform.

Fig. 3 shows the periodicity level results of the harmonic and inharmonic speech ($\alpha = 0.3$) calculated from 280 sentences in a corpus (Wong et al., 2007). Inharmonic speech has a significantly lower periodicity level ($t(279) = 91.1$, $p < 0.0001$, the paired-sample t-test) compared to harmonic speech, indicating the temporal periodicity was significantly decreased by jittering the harmonics with $\alpha = 0.3$.

2.2.2. Envelope modulation

To assess whether the jittering processing also affects envelope modulation, we compared the envelope modulation of both harmonic and inharmonic stimuli. Modulation spectra were computed using the front end of the mr-sEPSM speech intelligibility model (Jørgensen et al., 2013). If the modulation spectra were not changed by the jittering processing, then we can exclude the influence of envelope modulation on experimental conclusions.

The detailed calculation method for the modulation spectra of a given stimulus is as follows. The first stage of the intelligibility model uses a band-pass filterbank with 22 gammatone filters (Glasberg and Moore, 1990) spaced one-third octave apart, covering 63 Hz to 8 kHz. Temporal envelopes of outputs of the filterbank are extracted using the Hilbert transform followed by a first-order low-pass Butterworth filter with a 150-Hz cutoff frequency. Each channel envelope is then analyzed by filters in different modulation frequency ranges. In our study, all CI listeners use the ACE strategy in Cochlear® CI processors (Kiefer et al., 2001), so we replaced the Hilbert envelope with the ACE envelope. Specifically, the bandpass filterbank was implemented using a 128-point FFT and a Hanning window, grouping 65 bins into 22 frequency channels. The envelope of each channel was extracted by calculating the root of the summed bin powers.

Fig. 4(a)(b) displays the sample envelope modulation spectrograms from a sentence in its harmonic and inharmonic version. The sentence was "他切菜不小心切伤手指 (He carelessly cut his finger while cutting vegetables.)". It can be observed that the inharmonic speech had a similar envelope modulation spectrogram with its harmonic counterpart.

For simplicity, the modulation spectrograms were averaged across the 22 frequency channels for each sentence. Fig. 4(c) shows the average modulation spectrum across all 280 sentences from the MHINT corpus. We conducted a two-way repeated measures analysis of variance (rm-ANOVA), with harmonicity (Harmonic vs. Inharmonic) and modulation frequencies as independent variables, and modulation power as the dependent variable. Significant effects of modulation frequency ($F(7, 1953) = 5326$, $p < 0.0001$) on modulation power were observed, along with a significant interaction ($F(7, 1953) = 55.50$, $p < 0.0001$) between harmonicity and modulation frequency. However, no significant effect of harmonicity was observed ($F(1, 279) = 1.842$, $p = 0.1758$). These results indicate that while the jittering processing disrupts the harmonic structure and decreased periodicity correspondingly, it does not result in changes in the modulation spectra. Therefore, the inharmonic speech and the harmonic speech differed only in harmonicity, or periodicity. This finding is consistent with previous studies using the same jittering processing (Rajappa et al., 2023).

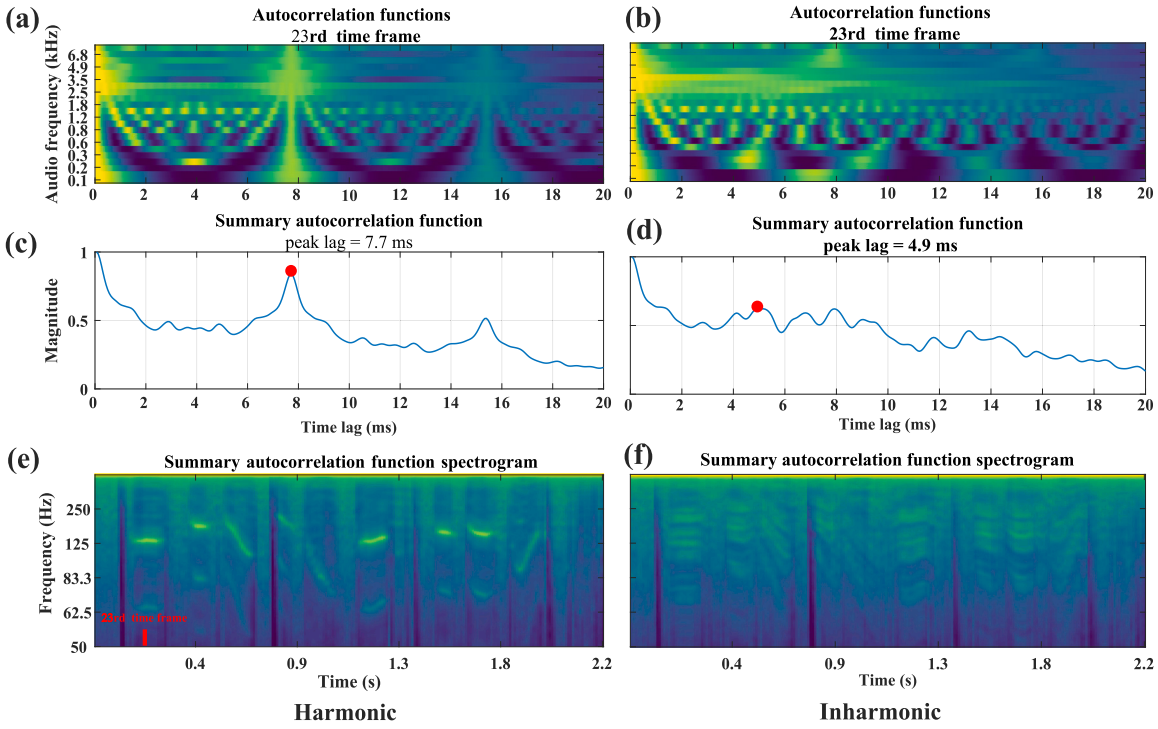


Fig. 2. Autocorrelation function of a tonal speech example under harmonic and inharmonic conditions. Top: Autocorrelation function for a random time frame (e.g., Time frame = 23). Middle: Summary autocorrelation function for the same frame. Bottom: Spectrograms of the summary autocorrelation function across all frames. The left panels are derived from a tonal sentence produced naturally (harmonic), while the right panels are derived from the same sentence with jittering by $\alpha = 0.3$ (inharmonic).

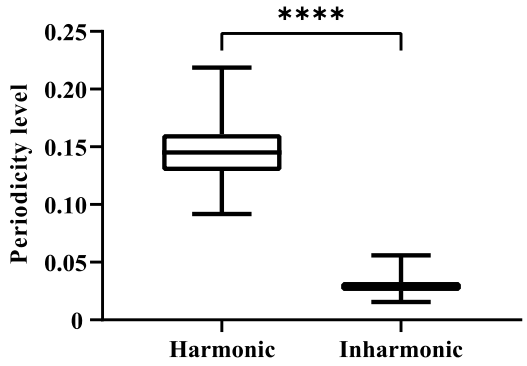


Fig. 3. Periodicity level results of the harmonic and inharmonic speech ($\alpha = 0.3$) calculated from 280 sentences in a corpus (Wong et al., 2007). Jittering with $\alpha = 0.3$ significantly decreased the periodicity level. The significance level $\alpha = 0.05$, when $p > 0.05$ means that the two results are not significantly different, expressed as *ns*, and the $p < 0.05$ is opposite. $*$ means $p < 0.05$, $**$ means $p < 0.01$, $***$ means $p < 0.001$, $****$ means $p < 0.0001$. The significance level and p-values in this study follow this convention.

2.3. Summary

The inharmonic speech was constructed by jittering the harmonics by $\alpha = 0.3$. Acoustic analysis show that, compared to harmonic speech, inharmonic speech has a decreased degree of periodicity, but the modulation spectra do not differ.

3. Experiments

To investigate the effects of harmonicity on speech perception in CI users, the experiments compare perception of harmonic and inharmonic speech in CI listeners and also in listeners with vocoder simulations. Two tasks were involved, speech perception in quiet and speech-on-speech perception.

3.1. Participants

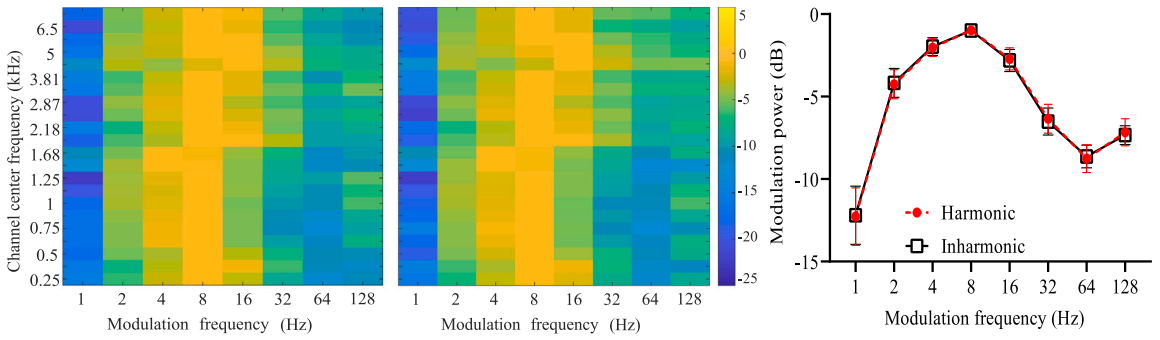
Two groups of adult CI listeners were involved in the experiments. One group ($N = 11$, listed in Table 1) participated in the speech-in-quiet task and the other group ($N = 20$, listed in Table 2) participated in the speech-on-speech task.

For comparisons, a total of $N = 49$ NH listeners (ages 19–27 years, mean = 23 years) were also involved. Thirteen of them participated in the speech-in-quiet task with vocoder simulations. The remaining 36 participated in the speech-on-speech task, with 20 for vocoder simulation (CI simulation group) and 16 for baseline without vocoder simulation (NH group). All the NH listeners (students recruited from Shenzhen University) self-reported normal hearing with no history of neurological or otologic pathology.

The study has been reviewed and approved by the Ethics Review Committee of Guangdong Provincial People's Hospital, with Ethics Number: KY2023-655-01. All participants signed written informed consent forms prior to the commencement of the experiments.

3.2. Stimuli

The speech-in-quiet task used sentence materials from two corpora: the Mandarin Hearing in Noise Test (MHINT) (Wong et al., 2007), spoken by an adult male talker, and the Mandarin Chinese Adaptation of AzBio (CMnBio) (Xi et al., 2022) spoken by four talkers (two males and two females). The MHINT corpus has 12 test lists and 2 practice lists, each containing 10 word sentences, and four test lists were randomly selected from the MHINT corpus for the quiet conditions, and these lists were not used in the speech-on-speech conditions. The CMnBio corpus is designed to minimize ceiling effect for Mandarin-speaking CI users (Xi et al., 2022). It has 20 sentence lists with a



(a) Harmonic (b) Inharmonic (c) The average envelope modulation power for each modulation filter

Fig. 4. Sample envelope modulation spectrograms from a sentence in its harmonic (a) and inharmonic version (b), and envelope modulation spectra from 280 sentences (c).

Table 1
CI participant demographics for speech in quiet.

ID	Gender	Age	Etiology	Device	CI experience (yr)
C2[a]	Female	23	Large vestibular aqueduct	CP910	6
C4	Female	25	Unknown	CP802	21
C5	Female	24	Unknown	Freedom	21
C6	Female	24	Congenital	Freedom	22
C7	Male	18	Congenital	CP1000	17
C12	Female	27	Unknown	CP900	4
C15	Female	36	Unknown	CP910	10
C21	Male	27	Drug-induced	CP1150	21
C22	Male	24	Drug-induced	CP802	18
C34	Female	44	Drug-induced	CP810	15
C37	Male	25	Unknown	CP1000	23
Mean		27			16.2

[a] This denotes bilaterally implanted users. For this study, bilaterally implanted users were tested unilaterally on their preferred ear to maintain consistency with the testing protocol for unilaterally implanted users.

Table 2
CI participant demographics for speech-on-speech condition.

ID	Gender	Age	Etiology	Device	CI experience (yr)
C1	Female	44	Drug-induced	CP810	15
C2[a]	Female	23	Large vestibular aqueduct	CP910	6
C3	Female	25	Drug-induced	CP910	19
C5	Female	24	Unknown	Freedom	21
C7	Male	18	Congenital	CP1000	17
C8	Male	28	Unknown	CP1150	19
C9	Male	24	Meningitis-induced	Freedom	5
C11	Female	31	Unknown	CP910	2
C13	Female	25	Unknown	CP802	21
C16	Male	51	Unknown	CP802	22
C19	Female	25	Large vestibular aqueduct	CP910	2
C20	Female	29	Unknown	CP910	2.5
C21	Male	27	Drug-induced	CP1150	21
C22	Male	24	Drug-induced	CP802	18
C24	Female	22	Congenital	Freedom	20
C27	Female	24	Congenital	CP910	21
C35[a]	Female	20	Large vestibular aqueduct	CP910	4
C37	Male	25	Unknown	CP1000	23
C42	Female	26	Large vestibular aqueduct	CP910	4
C46	Female	38	Unknown	ESPrit 3G	11
Mean		28			13.6

[a] This denotes bilaterally implanted users. For this study, bilaterally implanted users were tested unilaterally on their preferred ear to maintain consistency with the testing protocol for unilaterally implanted users.

variable number of words in a sentence, and 12 lists were used in this study. In the speech-on-speech task, both the target and masker sounds are from the MHINT corpus but from different lists. Sentences from the afore-mentioned corpora were used as the harmonic stimuli, and the inharmonic stimuli were generated by jittering the harmonic stimuli with $\alpha = 0.3$ as described in Section 2.1.

For vocoder simulations, the Pulsatile Gaussian-Enveloped Tones (GET) Vocoder (Meng et al., 2023) was used to process both harmonic (natural) and inharmonic speech, and the processed materials were presented to NH listeners. The GET system was used to convert the ACE output into vocoded sound, establishing a sequential process where ACE and GET worked together complementarily. Specifically, the electrical pulse from each electrode in the ACE output is converted into the corresponding envelope pulse through convolution with a Gaussian function. The envelope signal for each frequency band is then modulated with a standard sinusoidal wave to generate the simulated audio

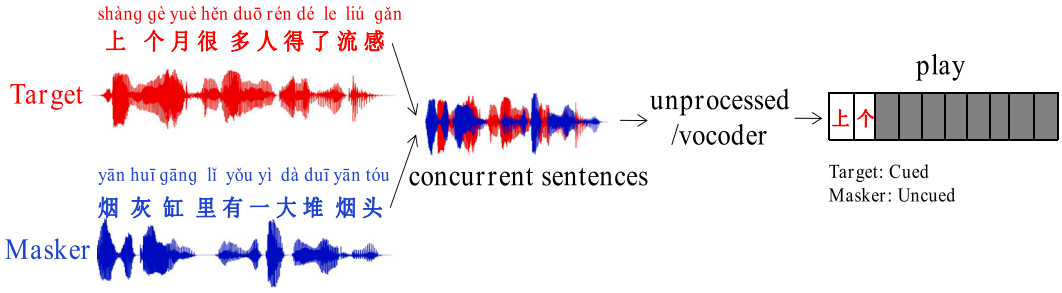


Fig. 5. An example to present the procedure of speech-on-speech condition. The target and masker were mixed and presented to listeners, the first two words of target sentence were cued on the computer.

signal. For more details on the process and electric channel interactions, please refer to Meng et al. (2023). Kong et al. (2023) has demonstrated that NH listeners with the GET vocoder yields comparable speech perception results to actual CI users.[1]

3.3. Procedure

The experiment was carried out in a soundproof room. Stimuli were presented through an audio interface and a loudspeaker located 1 m in front of the listener at a level of approximately 65 dB A-weighted sound pressure level (SPL).

In the speech-in-quiet task, speech recognition rate was measured using harmonic ($\alpha = 0$) and inharmonic ($\alpha = 0.3$) stimuli in NH listeners with vocoder simulations (the CI simulation group) and actual CI listeners (the CI group). Two lists were used for each condition, and the recognition rates were averaged. The CI simulation group was tested using the CMnBio corpus, and the CI group was tested using both the MHINT and CMnBio corpora. Listeners were asked to repeat the entire presented sentence as much as possible. Before the formal tests, a brief training session was conducted to familiarize listeners with the testing procedure, using one list of harmonic speech and one list of inharmonic speech.

In the speech-on-speech task, speech reception thresholds (SRTs) were measured using harmonic and inharmonic stimuli in three groups of listeners, namely, NH listeners without vocoder simulation (the NH group), NH listeners with vocoder simulation (the CI simulation group), and actual CI listeners (the CI group). Harmonic means both target (T) and masker (M) are harmonic (i.e., harmonic-on-harmonic) with $\alpha = 0$ (denoted by T0M0). Inharmonic means both target (T) and masker (M) are inharmonic (i.e., inharmonic-on-inharmonic) with $\alpha = 0.3$ (denoted by T0.3M0.3). Each condition was evaluated twice using two randomly selected lists from the MHINT corpus, and the results were averaged. A specific SNR was achieved by keeping the masker level constant and adjusting the speech level.

Fig. 5 presents the paradigm for the speech-on-speech condition. In each trial, the first two words of the target sentence were shown on a screen before stimuli presentation. Participants were instructed to focus on and repeat the target sentence that begins with these two cued words, while ignoring the masker sentence. A sentence was regarded as intelligible when at least five words (except for the two cued words) were repeated correctly. SRTs were measured with an adaptive procedure with the signal-to-noise ratio (SNR) initially set to 10 dB. The step size before the second reversal was 8 dB, followed by 4 dB before the fourth reversal, and 2 dB after that. The arithmetic mean of the SNRs for the last eight sentences was calculated as the SRT. Refer to Meng et al. (2016) for more details on the adaptive procedure.

[1] Examples of these stimuli are available for download from https://github.com/smylab/Harmonicity-and-CI/blob/main/Examples_of_the_vocoded_and_original_stimuli.zip.

3.4. Results

3.4.1. Behavioral results

For speech perception in quiet, as shown in Fig. 6, a consistent trend of decreased recognition rates for inharmonic speech compared to harmonic speech was observed for both the CI simulation group using CMnBio and the CI group using the two corpora. Specifically, the decrease was at a level of approximately 10 percentage points. The group mean recognition rate in the CI simulation group decreased from 72.3% (harmonic) to 61.6% (inharmonic), and this decrease is statistically significant ($t(12) = 4.27$, $p = 0.0011$, the paired-sample t-test). In the CI group, it significantly decreased from 70.2% (harmonic) to 61.2% (inharmonic) ($t(10) = 4.32, p = 0.0015$) with the CMnBio corpus, and from 94.5% (harmonic) to 85.2% (inharmonic) ($t(10) = 3.96, p = 0.0027$) with the MHINT corpus. Individual results show that, ceiling effects were observed for the MHINT corpus, whereas the CMnBio corpus reduced the occurrence of ceiling effects.

For speech-on-speech perception, inharmonic stimuli decreased performance compared to harmonic stimuli. Speech-on-speech results are shown in Fig. 7. For the NH group, paired t-tests revealed that participants presented with mixtures of T0M0 (mean SRT: –4.8 dB) had significantly better SRTs than those with T0.3M0.3 (–0.9 dB) ($t(15) = 4.316, p = 0.0006$). The SRTs are significantly higher for inharmonic speech-on-speech tests than for harmonic speech-on-speech tests, for both CI listener groups (the CI simulation and CI groups). We conducted separate paired t-tests comparing T0M0 and T0.3M0.3 as the independent variable and SRTs as the dependent variable. The CI simulation group showed a difference of t(19) = 2.854,p = 0.0106, while for the CI group it was t(19) = 3.69,p = 0.0017. Specifically, the mean SRTs were 8 dB and 10.8 dB for the CI simulation group, and 8.8 dB and 11.6 dB for the CI group under the T0M0 and T0.3M0.3 conditions, respectively. Furthermore, the group comparison revealed that the SRTs of the NH group were mostly negative, while the SRTs of both the simulated and CI groups were positive, primarily ranging from 5 to 15 dB. A two-way measures ANOVA was conducted between the CI simulation and CI groups, results showed there were no significant difference ($F(1, 1) = 71.8, p = 0.07$). Additionally, we observed different levels of SRT variability in the NH group under the T0M0 and T0.3M0.3 conditions, which may suggest substantial individual differences in the utilization of these auditory cues.

3.4.2. Psychometric functions fitting

To examine the effect of harmonicity on Mandarin speech perception in both NH and CI listeners, psychometric functions were fitted to the data from both groups, aiming to derive generalizable conclusions. The data consist of two groups: (1) NH listeners without vocoder simulation (i.e., NH group, shown on the left side of Fig. 7), and (2) actual CI listeners (i.e., CI group, shown on the right side of Fig. 7). In Section 3.4.1, it is reported that there was no significant difference

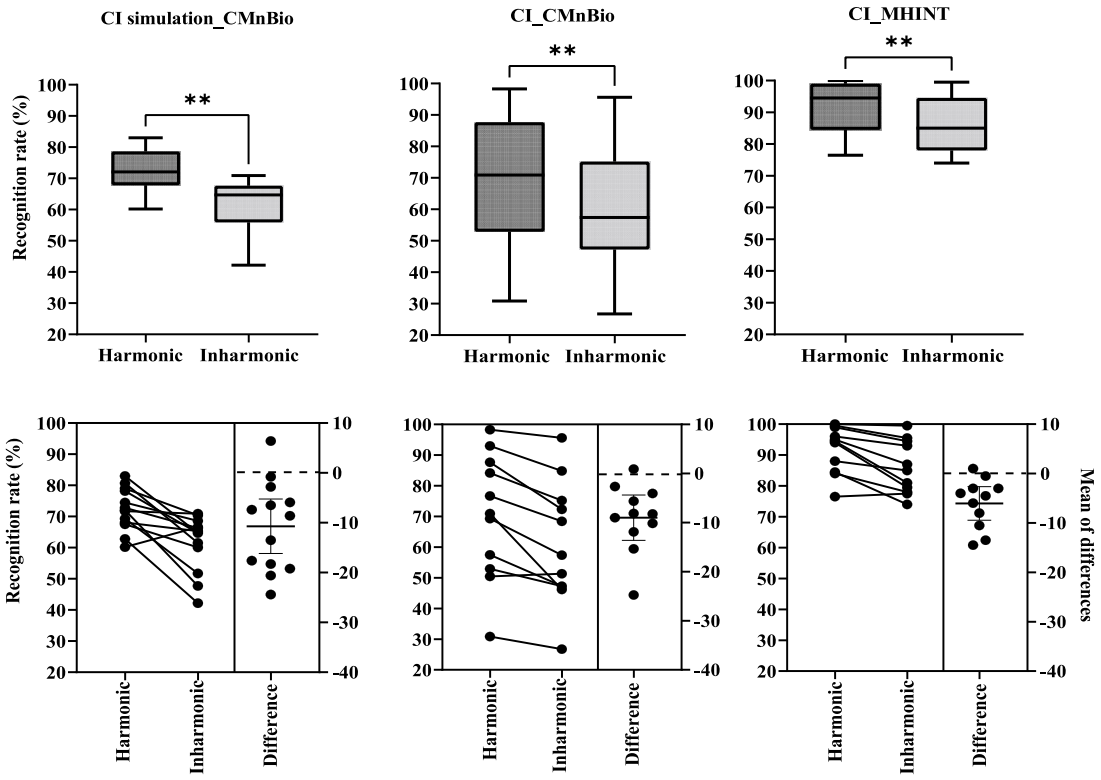


Fig. 6. Group (top) and individual (bottom) recognition rate of harmonic and inharmonic speech in quiet condition. Difference: the recognition rate difference between harmonic and inharmonic conditions. CI simulation_CMnBio: CI simulation test with CMnBio corpus. CI_CMnBio and CI_MHINT: CI group with CMnBio and MHINT corpus, respectively. Error bars indicate standard deviations and the asterisks above denote the statistical significance of frequency jitter's effect on the recognition rate.

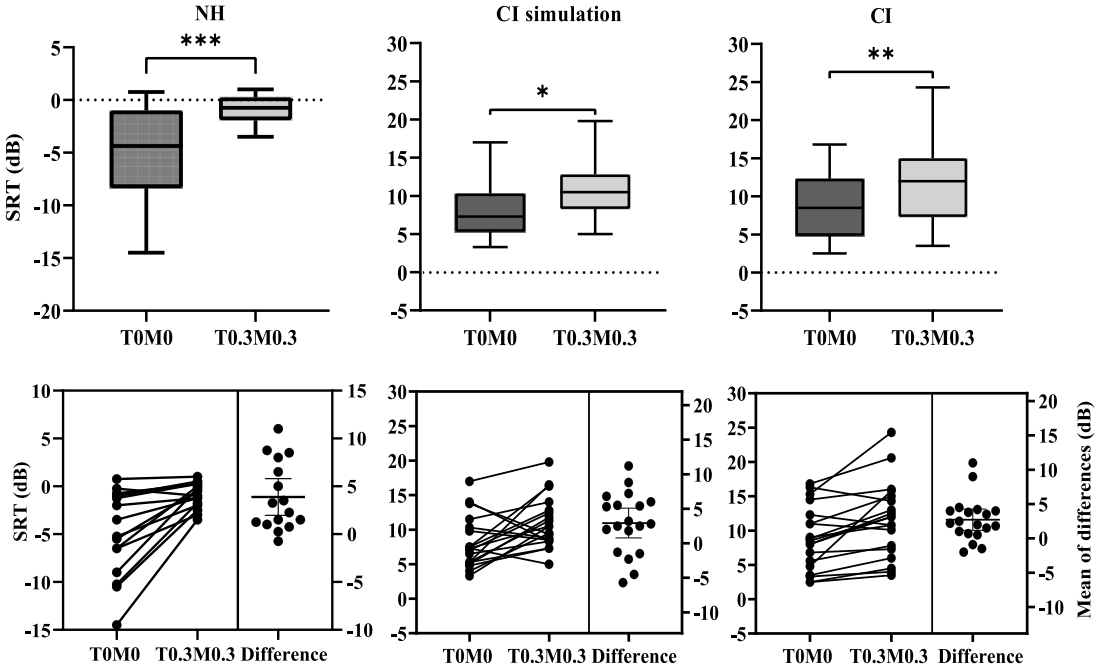


Fig. 7. Boxplots of the SRTs for NH, CI simulation, and CI groups obtained in speech-on-speech condition. Error bars indicate the standard deviations, and asterisks above mark the statistical significance of frequency jitter's effect on the SRT.

between the CI group and the CI simulation group (NH listeners with vocoder simulation), suggesting that both groups reflect the speech perception abilities of actual CI users. Therefore, to simplify the fitting results, speech intelligibility functions were derived only from the NH and CI group data tested with the MHINT corpus, excluding the CI simulation group data. The fitted curves are shown in Fig. 8. The following equation was used for fitting.

$$\Psi(x; s, t, \lambda, \gamma) = \gamma + (1 - \lambda - \gamma)\frac{1}{1 + e^{-s(x-t)}} \tag{3}$$

where Ψ denotes the speech intelligibility, γ is the chance level, λ is the lapse rate, t denotes the threshold, and s is the slope of the

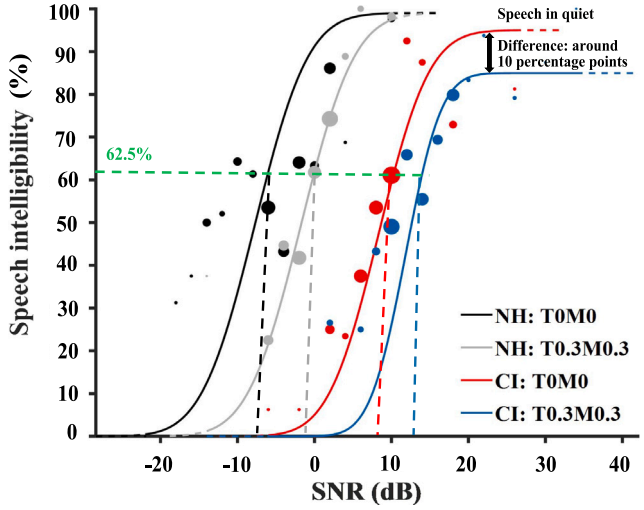


Fig. 8. Fitted psychometric functions for the MHINT corpus in both quiet and speech-on-speech conditions for NH group and CI group. 62.5% (5 out of 8 words) was the threshold for the adaptive SRT measurement in the speech-on-speech experiment.

psychometric curve at the threshold point. The fitting was conducted using the Psignifit toolbox developed by the Wichmann lab (Schütt et al., 2016). λ was set as a free parameter for NH listeners. For CI listeners, it was set at a value of 0.05 and 0.15 for harmonic and inharmonic stimuli, respectively. These values were based on the recognition rates in quiet conditions, where mean recognition rates (see Fig. 6, CI_MHINT group) are 94.5% in harmonic conditions and 85.2% in inharmonic conditions, corresponding to λ values of 0.05 (1-0.95) and 0.15 (1-0.85), respectively.

Several observations can be made from the curves in Fig. 8, which reflect differences in speech perception across various conditions between NH and CI listeners.

- For speech perception in quiet, harmonicity did not affect NH listeners' performance but decreased that of the CI listeners', as indicated by the overlapped upper asymptotes of the two curves of NH listeners and the vertical gap between the two upper asymptotes of CI listeners. To ensure rigor, we conducted an additional speech perception experiment under quiet conditions with 5 NH listeners without vocoder (the NH group) using the MHINT and CMnBio corpora. The results (not shown here) were consistent with the predictions under both harmonic and inharmonic conditions.
- For speech-on-speech perception, both NH and CI listeners performed worse with inharmonic stimuli compared to harmonic stimuli, which could be indicated by the horizontal gaps (or SRT differences) between the two NH curves as well as between the two CI curves.
- While NH listeners can understand inharmonic speech even in negative SNRs, CI listeners can only understand speech in positive SNRs. In positive SNRs, CI listeners' speech-on-speech perception may be dominated by the inherent advantage of harmonic over inharmonic stimuli in quiet environments.

4. Discussions

4.1. Speech perception in quiet

The main contribution of this study is the finding that harmonicity significantly impacts CI listeners' Mandarin speech perception in quiet. The recognition rate for inharmonic sentences was approximately 10 percentage points lower than that for harmonic sentences in CI listeners ($p < 0.05$). In contrast, previous studies have shown that a lack of harmonicity does not affect NH listeners' English speech perception in quiet (Popham et al., 2018). This distinction is also supported by the overlapping upper asymptotes of the fitted NH curves for Mandarin speech in Fig. 8, which align with the results of an additional experiment conducted in our study with 5 NH listeners (detailed results not included in this manuscript).

The differences in the effects of harmonicity on speech perception in quiet can be explained by the various acoustic cues utilized by CI and NH listeners. Fine temporal and spectral cues, like resolved harmonics and temporal fine structure, are accessible to NH listeners. However, the limited spectral resolution of CIs highlights the critical role of temporal cues for CI users. Unlike CI users, NH listeners are still able to deduce sufficient information from other temporal and spectral cues to facilitate speech understanding, even when the temporal periodicity of the inharmonic speech has diminished. Chen et al. (2014) found that the loss of tonal contour information alone, while preserving the harmonic structure to some degree, has no effect on Mandarin sentence intelligibility in quiet. Furthermore, McPherson and McDermott (2018) reported that tone identification was comparable for harmonic and inharmonic Mandarin speech but decreased substantially for whispered speech. These findings suggest that NH listeners evidently track the frequency contours of stimuli, regardless of whether the frequencies are harmonic or inharmonic. However, CI listeners, lacking sufficient temporal and spectral cues, may struggle to compensate for a decline in periodicity, which could be detrimental to their speech perception.

It is important to note that the effects of harmonicity might vary across studies depending on whether tonal or non-tonal speech materials are used. Tonal languages like Mandarin rely heavily on pitch contour cues, which could influence the observed effects of harmonicity. In contrast, studies using non-tonal languages might place less emphasis on such cues, potentially leading to different outcomes. This distinction underscores the need for careful consideration of language context when interpreting harmonicity effects.

4.2. Speech-on-speech perception

Speech-on-speech perception is more complex than speech-in-quiet perception as it involves masking release. In our speech-on-speech experiment, both NH listeners (NH group) and CI users (CI simulation and CI groups) performed worse with the inharmonic stimuli compared to the harmonic stimuli. The results for NH listeners are consistent with the study by Popham et al. (2018), which suggests that harmonicity contributes to the grouping and streaming of non-tonal natural speech. McPherson et al. (2022) indicated that the benefits of harmonicity on the detection of non-tonal target sounds occur regardless of the harmonicity of the competing speaker. Wu (2019) suggested that unnatural pitch contours reduce the release effect of Mandarin speech in masking environments, meaning that if the target speech pitch becomes unnatural, its masking release ability weakens. Steinmetzger and Rosen (2023) and Steinmetzger et al. (2019) found that NH listeners hardly benefitted from the periodicity (i.e., harmonicity) in non-tonal target speech (Steinmetzger et al., 2019), and their latest research (Steinmetzger and Rosen, 2023) reported that regular spacing of spectral components rather than masker's harmonicity plays a role. It should be noted that the maskers in their study were unintelligible, so their findings may not extend to speech-on-speech masking scenarios. Future research, using models optimized for analyzing natural auditory scenes, may help elucidate the basis of these different effects.

For CI users, neither regular spacing of spectral components nor the harmonic "cancellation" model (the idea that the auditory system 'cancels' harmonic masking sounds in order to identify concurrent target sounds, rather than 'enhancing' harmonic targets themselves) (de Cheveigné, 2021, 1993) significantly impacts task performance because they cannot decompose individual harmonic components. Steinmetzger and Rosen (2018) demonstrated that CI users experience significant

masking release from masker periodicity, as implemented through harmonic complex tones with pitch-related periodic envelope modulations, compared to aperiodic noise. This release arises from temporal envelope cues but not from slow envelope modulations. Our study further suggests that CI users benefit substantially from target harmonicity when the target sentence is in Mandarin, a tonal language. As mentioned in Section 3.4.2, CI users can successfully perceive target speech at positive SNRs. This makes it difficult to determine whether harmonicity aids masking release for CI users, since the observed results under speech-on-speech conditions may simply reflect the inherent differences in intelligibility between harmonic and inharmonic sentences in quiet conditions. Future research should design more targeted experiments to further investigate this issue.

5. Conclusions

This study demonstrates that NH listeners (NH group) and CI users (CI simulation and CI groups) rely on different acoustic cues for speech recognition. We primarily associate the role of harmonicity with temporal cues, as harmonic components form periodic patterns in the temporal domain, which are crucial for CI users' speech perception in tonal languages like Mandarin. In quiet environments, harmonicity does not affect speech recognition in NH listeners but decreases performance in CI users for Mandarin speech, likely due to its differential impact on Mandarin tone perception. While the lack of harmonicity does not affect tone perception in NH listeners, it may impair tone perception in CI users due to inharmonic manipulation disrupting periodicity, which is associated with lexical tone perception. In speech-on-speech conditions, both NH listeners (NH group) and CI users (CI simulation and CI groups) perform worse with inharmonic stimuli. NH listeners can understand inharmonic speech even at negative SNRs, suggesting that harmonicity aids in masking release. In contrast, CI users require positive SNRs (often above 5 dB) to understand speech, indicating that their performance depends more on the intelligibility of the target speech than on benefits from masking release. Future research should further explore the effects of harmonicity on tone perception, as well as how target and background speech harmonicity influences masking release in CI users. Additionally, improving the transmission of harmonicity cues in cochlear implant coding strategies could potentially enhance speech perception in tonal languages.

CRediT authorship contribution statement

Mingyue Shi: Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation. **Qinglin Meng:** Writing – review & editing, Validation, Supervision, Resources, Methodology. **Huali Zhou:** Writing – review & editing, Software, Methodology. **Jiawen Li:** Validation, Investigation. **Yefei Mo:** Validation, Investigation. **Nengheng Zheng:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of generative AI

During the preparation of this work, we utilized ChatGPT-4 for spell-checking and grammar review. Following the use of this tool/service, we conducted a thorough review and made necessary edits, taking full responsibility for the content of the publication.

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.

Acknowledgments

The authors would like to thank all the research volunteers that generously donated their time to participate in this study. This work was supported by the Guangdong Basic and Applied Basic Research Foundation Grant (2022A1515011361, 2024A1515012585), the Shenzhen Fundamental Research Program (20220809191805001), China Disabled Persons' Federation (CDPF) Hearing and Speech Program (2024CDPFHS-20), and the China National Natural Science Foundation (12374448). Additionally, we express our gratitude to Dr. Xin Xi for providing the CMnBio corpus, and to Dr. Kurt Steinmetzger and Dr. Stuart Rosen for their support on the calculation of the periodicity levels.

Data availability

Data will be made available on request.

References

Chen, F., Wong, L.L., Hu, Y., 2014. Effects of lexical tone contour on Mandarin sentence intelligibility. J. Speech Lang. Hear. Res. 146 (2), EL99–EL105.
Cooper, H.R., Roberts, B., 2007. Auditory stream segregation of tone sequences in cochlear implant listeners. Hear. Res. 225 (1–2), 11–24.
de Cheveigné, A., 1993. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. J. Acoust. Soc. Am. 93 (6), 3271–3290.
de Cheveigné, A., 2021. Harmonic cancellation—A fundamental of auditory scene analysis. Trends Hear. 25, 23312165211041422.
de Cheveigné, A., McAdams, S., Marin, C.M., 1997. Concurrent vowel identification. II. Effects of phase, harmonicity, and task. J. Acoust. Soc. Am. 101 (5), 2848–2856.
Glasberg, B.R., Moore, B.C.J., 1990. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 47 (1–2), 103–138.
Goldsworthy, R.L., 2019. Temporal envelope cues and simulations of cochlear implant signal processing. Speech Commun. 109, 24–33.
Jørgensen, S., Ewert, S.D., Dau, T., 2013. A multi-resolution envelope-power based model for speech intelligibility. J. Acoust. Soc. Am. 134 (1), 436–446.
Kawahara, H., Morise, M., 2011. Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. Sadhana 36 (5), 713–727.
Kiefer, J., Hohl, S., Stürzebecher, E., Pfennigdorff, T., Gstöttner, W., 2001. Comparison of speech recognition with different speech coding strategies SPEAK, CIS, and ACE and their relationship to telemetric measures of compound action potentials in the nucleus CI 24m cochlear implant system.. Audiology 40 (1), 32–42.
Kong, F., Zhou, H., Mo, Y., Shi, M., Meng, Q., Zheng, N., 2023. Comparable encoding, comparable perceptual pattern: Acoustic and electric hearing. IEEE Trans. Neural Syst. Rehabil. Eng..
Lentz, B., Völter, C., Martin, R., 2022. Spectral sparsification of speech signals and its interaction with top-down mechanisms in adult cochlear implant users. Speech Commun. 144, 67–74.
McPherson, M.J., 2022. The Perception of Harmonic Sounds. Harvard University.
McPherson, M.J., Grace, R.C., McDermott, J.H., 2022. Harmonicity aids hearing in noise. Atten. Percept. Psychophys. 84 (3), 1016–1042.
McPherson, M.J., McDermott, J.H., 2018. Diversity in pitch perception revealed by task dependence. Nat. Hum. Behav. 2 (1), 52–66.
Meddis, R., Hewitt, M.J., 1991. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. J. Acoust. Soc. Am. 89 (6), 2866–2882.
Meddis, R., O'Mard, L., 1997. A unitary model of pitch perception. J. Acoust. Soc. Am. 102 (3), 1811–1820.
Meng, Q., Zheng, N., Li, X., 2016. Mandarin speech-in-noise and tone recognition using vocoder simulations of the temporal limits encoder for cochlear implants. J. Acoust. Soc. Am. 139 (1), 301–310.
Meng, Q., Zhou, H., Lu, T., Zeng, F.-G., 2023. Pulsatile Gaussian-enveloped tones (GET) for cochlear-implant simulation. Appl. Acoust. 208, 109386.
Micheyl, C., Oxenham, A.J., 2010. Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. Hear. Res. 266 (1–2), 36–51.
Oxenham, A.J., 2008. Pitch perception and auditory stream segregation: implications for hearing loss and cochlear implants. Trends Amplif. 12 (4), 316–331.
Plack, C.J., 2018. The Sense of Hearing. Routledge.
Popham, S., Boebinger, D., Ellis, D.P., Kawahara, H., McDermott, J.H., 2018. Inharmonic speech reveals the role of harmonicity in the cocktail party problem. Nat. Commun. 9 (1), 1–13.
Qin, M.K., Oxenham, A.J., 2003. Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. J. Acoust. Soc. Am. 114 (1), 446–454.

Qin, M.K., Oxenham, A.J., 2005. Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification. Ear Hear. 26 (5), 451–460.
Rajappa, N., Guest, D.R., Oxenham, A.J., 2023. Benefits of harmonicity for hearing in noise are limited to detection and pitch-related discrimination tasks. Biology 12 (12), 1522.
Schütt, H.H., Harmeling, S., Macke, J.H., Wichmann, F.A., 2016. Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. Vis. Res. 122, 105–123.
Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270 (5234), 303–304.
Spitzer, S., Liss, J., Spahr, T., Dorman, M., Lansford, K., 2009. The use of fundamental frequency for lexical segmentation in listeners with cochlear implants. J. Acoust. Soc. Am. 125 (6), EL236–EL241.
Steinmetzger, K., Rosen, S., 2018. The role of envelope periodicity in the perception of masked speech with simulated and real cochlear implants. J. Acoust. Soc. Am. 144 (2), 885–896.
Steinmetzger, K., Rosen, S., 2023. No evidence for a benefit from masker harmonicity in the perception of speech in noise. J. Acoust. Soc. Am. 153 (2), 1064–1072.
Steinmetzger, K., Zaar, J., Relano-Iborra, H., Rosen, S., Dau, T., 2019. Predicting the effects of periodicity on the intelligibility of masked speech: An evaluation of different modelling approaches and their limitations. J. Acoust. Soc. Am. 146 (4), 2562–2576.
Stickney, G.S., Assmann, P.F., Chang, J., Zeng, F.-G., 2007. Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences. J. Acoust. Soc. Am. 122 (2), 1069–1078.
Wong, L.L., Soli, S.D., Liu, S., Han, N., Huang, M.-W., 2007. Development of the Mandarin hearing in noise test (MHINT). Ear Hear. 28 (2), 70S–74S.
Wu, M., 2019. Effect of F0 contour on perception of Mandarin Chinese speech against masking. PLoS One 14 (1), e0209976.
Xi, X., Wang, Y., Shi, Y., Gao, R., Li, S., Qiu, X., Wang, Q., Xu, L., 2022. Development and validation of a Mandarin Chinese adaptation of azbio sentence test (CMnBio). Trends Hear..