frontiers in Neuroscience

ORIGINAL RESEARCH
published: 02 December 2021
doi: 10.3389/fnins.2021.744959

Relative Weights of Temporal Envelope Cues in Different Frequency Regions for Mandarin Vowel, Consonant, and Lexical Tone Recognition

Zhong Zheng[1,2], Keyi Li[3], Gang Feng[4], Yang Guo[5], Yinan Li[1,2], Lili Xiao[1,2], Chengqi Liu[1,2], Shouhuan He[6], Zhen Zhang[1,2], Di Qian[7]* and Yanmei Feng[1,2]**

[1] Department of Otolaryngology-Head and Neck Surgery, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China, [2] Shanghai Key Laboratory of Sleep Disordered Breathing, Shanghai, China, [3] Sydney Institute of Language and Commerce, Shanghai University, Shanghai, China, [4] Department of Graduate, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, China, [5] Ear, Nose, and Throat Institute and Otorhinolaryngology Department, Eye and ENT Hospital of Fudan University, Shanghai, China, [6] Department of Otolaryngology, Qingpu Branch of Zhongshan Hospital Affiliated to Fudan University, Shanghai, China, [7] Department of Otolaryngology, Shenzhen Longhua District People's Hospital, Shenzhen, China

OPEN ACCESS

Edited by:
Lin Chen,
University of Science and Technology of China, China

Reviewed by:
Yi-Wen Liu,
National Tsing Hua University, Taiwan
Chenghui Jiang,
Nanjing Medical University, China

***Correspondence:**
Zhen Zhang
zhangzhen1994s@outlook.com
Di Qian
skeayqd@sina.com
Yanmei Feng
ymfeng@sjtu.edu.cn

Specialty section:
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience

Received: *22 July 2021*
Accepted: *15 November 2021*
Published: *02 December 2021*

Citation:
Zheng Z, Li K, Feng G, Guo Y, Li Y, Xiao L, Liu C, He S, Zhang Z, Qian D and Feng Y (2021) Relative Weights of Temporal Envelope Cues in Different Frequency Regions for Mandarin Vowel, Consonant, and Lexical Tone Recognition. Front. Neurosci. 15:744959. doi: 10.3389/fnins.2021.744959

Objectives: Mandarin-speaking users of cochlear implants (CI) perform poorer than their English counterpart. This may be because present CI speech coding schemes are largely based on English. This study aims to evaluate the relative contributions of temporal envelope (E) cues to Mandarin phoneme (including vowel, and consonant) and lexical tone recognition to provide information for speech coding schemes specific to Mandarin.

Design: Eleven normal hearing subjects were studied using acoustic temporal E cues that were extracted from 30 continuous frequency bands between 80 and 7,562 Hz using the Hilbert transform and divided into five frequency regions. Percent-correct recognition scores were obtained with acoustic E cues presented in three, four, and five frequency regions and their relative weights calculated using the least-square approach.

Results: For stimuli with three, four, and five frequency regions, percent-correct scores for vowel recognition using E cues were 50.43–84.82%, 76.27–95.24%, and 96.58%, respectively; for consonant recognition 35.49–63.77%, 67.75–78.87%, and 87.87%; for lexical tone recognition 60.80–97.15%, 73.16–96.87%, and 96.73%. For frequency region 1 to frequency region 5, the mean weights in vowel recognition were 0.17, 0.31, 0.22, 0.18, and 0.12, respectively; in consonant recognition 0.10, 0.16, 0.18, 0.23, and 0.33; in lexical tone recognition 0.38, 0.18, 0.14, 0.16, and 0.14.

Conclusion: Regions that contributed most for vowel recognition was Region 2 (502–1,022 Hz) that contains first formant (*F*1) information; Region 5 (3,856–7,562 Hz) contributed most to consonant recognition; Region 1 (80–502 Hz) that contains fundamental frequency (F0) information contributed most to lexical tone recognition.

Keywords: temporal envelope cues, frequency region, Mandarin, vowel, consonant, tone

INTRODUCTION

Hearing loss is a common sensory disorder and has become an important global health problem due to the increasing prevalence and its negative impact on quality of life. World Health Organization [WHO] (2020) estimates that 466 million people suffer from hearing loss, with sensorineural hearing loss (SNHL) being the most common. Cochlear implant (CI) is currently the only effective method for patients with severe-to-profound SNHL (Zeng, 2004). Plenty of past research has been conducted to figure out the best strategies for encoding speech. Roy et al. (2015) programmed with either fine structure processing or high-definition continuous interleaved sampling strategy for CI users, and found that fine structure processing strategy offers better musical sound quality discrimination for CI users with respect to fundamental frequency perception. Tabibi et al. (2020) implemented a bio-inspired coding strategy for better representation of spectral and temporal information with 11 CI users, and significantly better performance was observed for bio-inspired coding strategy compared to the advanced combination encoder strategy. Recently, there are many studies for tonal language pitch encoding. Temporal limits encoder, optimized pitch, and language strategy has recently been proposed that can provide a significant benefit to perception of speech intonation (Meng et al., 2016; Vandali et al., 2019). The mainstream CI speech processing strategies, such as advanced combination encoder (Psarros et al., 2002), SPEAK (Skinner et al., 2002), and n-of-m (Ziese et al., 2000; Buechner et al., 2009) are based on the continuous interleaved sampling strategy (Wilson et al., 1991; Boëx et al., 1996). For the continuous interleaved sampling speech processing strategy, the electrode array is successively spaced with a single stimulus, that is, only one electrode is emitting the stimulus current at a time, and the interference and diffusion of the stimulus current between two electrodes are prevented by alternating stimulation (Zeng et al., 2008). Although contemporary CI has up to 22 intracochlear electrodes, the capacity of patients to use multiple channels typically asymptotes at around 8 channels or less (Pfingst et al., 2011; Macherey and Carlyon, 2014). Therefore, even as the most successful neural implants in the world, there is still much to be studied and improved in signal processing strategies.

Speech acoustic signals can be regarded as the temporal envelope (E) cues with slow change and the temporal fine structure (TFS) information with fast change based on the Hilbert transform (Kong and Zeng, 2006). The TFS information is the pre-dominant cues for lexical tone perception in NH listeners (Xu and Pfingst, 2003) but in hearing-impaired listeners and in noise environment, envelope cue plays an increasingly important role for lexical tone perception (Wang S. et al., 2011; Qi et al., 2017). The temporal E cues represents the amplitude of the waveform changing with time phase, which usually includes the duration information and amplitude E cues of the speech signal (Kong and Zeng, 2006). Perceptual research has shown that E cues are important for speech perception in quiet conditions (Smith et al., 2002; Xu and Pfingst, 2003). Different frequency regions of speech signals contain different information with varying functions, making it necessary to evaluate the relative importance of temporal information with different frequency regions in speech recognition. Past research methods on the role of temporal information in different frequency regions for speech recognition include removing a specific spectral information (Shannon et al., 2002), correlation analysis (Apoux and Bacon, 2004), lowpass- and highpass- filtration (Ardoint and Lorenzi, 2010), and band-pass filtration (Ardoint et al., 2011).

Previous research was mostly based on English, a non-tonal language. Mandarin, a tonal and most common spoken language in the world is significantly different from English. Mandarin includes 24 finals, 23 consonants, and 4 lexical tones. The 24 finals include 6 monophthongs, 9 diphthongs, and triphthongs, and 9 nasal finals. The 23 consonants always occur as "initials." Phonemes that include vowels and consonants are important signals for the auditory system because they have great contribution to speech intelligibility across languages (Kewley-Port et al., 2007). The four lexical tones include Lexical tone 1- (high-level), Lexical tone 2/(rising), Lexical tone 3 v (falling-rising), and Lexical tone 4 \(falling). In Mandarin, the same words with different lexical tones can represent many different meanings (Nissen et al., 2005). Previous studies have shown that compared with normal-hearing listeners, Mandarin-speaking CI users have shown poor performance in lexical tone recognition (Wei et al., 2004; Wang W. et al., 2011). One of the most important reasons is that most Chinese CI wearers have imported devices where the language processing strategy is calibrated toward non-tonal languages. Our previous studies have shown that the acoustic temporal E cues in frequency regions 1 (80–502 Hz) and 3 (1,022–1,913 Hz) significantly contributed to Mandarin sentence recognition in quiet (Guo et al., 2017). Given that speech perception involves both bottom-up and top-down processes, sentence recognition are heavily influenced by phonemic, lexical tone, and context in top-down condition (Hickok and Poeppel, 2007). This was the basis for our investigation into the different contributions of frequency regions in Mandarin phonemic and lexical tone recognition.

MATERIALS AND METHODS

Subjects

A group of 11 listeners (5 males, 6 females) from graduates of Shanghai Jiao Tong University were recruited in this study. Their ages ranged from 22 to 27 (average = 24.6) years with no reported history of ear disease or hearing difficulty. They were all native Mandarin Chinese speakers with normal audiometric thresholds (<20 dB HL), bilaterally, at frequencies between 0.25 and 8 kHz. Pure-tone audiometric thresholds were recorded using a GSI-61 audiometer (Grason-Stadler, Madison, WI, United States) with standard audiometric procedures. All subjects had no preceding exposure to the speech materials. Before the experiment, all subjects had signed a consent form and were compensated hourly. All procedures performed in studies involving human participants were approved and in accordance with the Ethics Committee of the Sixth People's Hospital affiliated to Shanghai Jiao Tong University (ChiCTR-ROC-17013460) and with the 1964 Declaration of Helsinki and its later amendments.

Signal Processing

The speech test program named Angel Sound[1] developed by Qian-Jie Fu at the House Ear Institute (Los Angeles, CA, United States) was used for Mandarin phoneme and lexical tone tests (Wu et al., 2007). All speech materials were sourced from the language database of University of Science and Technology of China, which includes the phonemes and lexical tones most frequently used in Mandarin. All the materials were recorded by one male and one female native Mandarin speaker. All speech stimuli were sampled at a 22-kHz sampling rate, without high-frequency pre-emphasis. The test ensures that only one phoneme is different. For example, to test for vowel, combinations of the same consonant and lexical tone that carry the most vowel options were selected. For lexical tone, given the maximum possibility is four, phoneme combinations fewer than four lexical tones were excluded (Wu et al., 2007). Lexical tone duration was normalized for lexical tone tokens to minimize bias on tonal perception (Jing et al., 2017). The speech materials were filtered into 30 contiguous frequency bands using zero-phase, third-order Butterworth filters (18 dB/oct slopes), ranging from 80 to 7,562 Hz (Li et al., 2016; Guo et al., 2017; Zheng et al., 2021). Each frequency band was an equivalent rectangular bandwidth for normal people, which simulates the frequency selection of normal auditory system (Glasberg and Moore, 1990). E information was extracted from each band using the Hilbert decomposition and low-pass filter at 64 Hz using third-order Butterworth filters. Then E was used to modulate the amplitude of a white noise. The envelope-modulated noise was bandpass-filtered using the same filter parameters as before. This study focuses on the parameters used in the present CI strategy in low frequency (<500 Hz), medium low frequency (500–1,000 Hz), medium frequency (1,000–2,000 Hz), medium high frequency (2,000–4,000 Hz), and high frequency (4,000–8,000 Hz) bands. Given the cut-off frequency of each frequency band is close to 500, 1,000, 2,000, 4,000, and 8,000 Hz, the modulated noise bands were summed across frequency bands to produce the frequency regions of acoustic E cues as follows: Bands 1–8, 9–13, 14–18, 19–24, and 25–30 were summed to form Frequency Regions 1–5, respectively (**Table 1**). To prevent subjects from using the E cues of the adjacent boundary band (Warren et al., 2004; Li et al., 2015), the frequency region containing the E cues was combined with complementary frequency regions containing noise masker that was presented at a speech-to-noise ratio of +16 dB. The speech-to-noise ratio was determined prior to signal processing using a full range of speech and noise stimuli. Masking noise was low-pass and high-pass filtered so that the final long-term power spectrum did not overlap the processed speech signals as the previous study (Ardoint et al., 2011). To investigate the role of different frequency regions for Mandarin phoneme and lexical tone recognition, the E cues from three frequency regions (10 conditions including "Region 123," "Region 124," "Region 125," "Region 134," "Region 135," "Region 145," "Region 234," "Region 235," "Region 245," and "Region 345"), four frequency regions (five conditions including "Region 1234," "Region 1345," "Region 1245," "Region 1235," and "Region 2345") and five frequency regions (one condition, "Region 12345") were presented to subjects. For example, the condition of "Region 123" meant the stimulus presented to the subject contained the E cues of frequency regions 1, 2, and 3 with noise of the remaining frequency regions 4 and 5. Similarly, in the test condition of "Region 124," the stimulus sound contains the E cues of frequency regions 1, 2, and 4 while other frequency bands (band 3 and 5) are white noise. In the test of full band region "Region 12345," the stimulus sound contains the E cues information of all frequency bands, and there is no other noise.

TABLE 1 | Cut-off frequency for extracting temporal envelope information in different frequency regions.

Frequency regions	Bands	Lower frequency (Hz)	Upper frequency (Hz)
1	1	80	115
	2	115	154
	3	154	198
	4	198	246
	5	246	300
	6	300	360
	7	360	427
	8	427	502
2	9	502	585
	10	585	677
	11	677	780
	12	780	894
	13	894	1,022
3	14	1,022	1,164
	15	1,164	1,322
	16	1,322	1,499
	17	1,499	1,695
	18	1,695	1,913
4	19	1,913	2,157
	20	2,157	2,428
	21	2,428	2,729
	22	2,729	3,066
	23	3,066	3,440
	24	3,440	3,856
5	25	3,856	4,321
	26	4,321	4,837
	27	4,837	5,413
	28	5,413	6,054
	29	6,054	6,767
	30	6,767	7,562

Test Procedure

None of the subjects had participated in the perception experiments testing acoustic temporal E cues before. The experiments were conducted in a double-walled, soundproof room. All test stimuli were delivered through Sennheiser HD205 II circumaural headphones. The stimuli were determined according to the most comfortable level of the subjects, generally around 65 dB SPL.

[1] http://angelsound.tigerspeech.com/

Before the formal test, ∼30 min of practice were provided. The speech material was presented under "full Region" conditions initially, and then presented in the same way as the test condition stimulus. Feedback was given during the practice. To familiarize the subjects with the test material, they can repeatedly listen to a word indefinitely and move on after they feel they have reached a stable state.

In the formal test, we randomly selected test sounds from different conditions and allowed subjects to hear the same test sound multiple times. Subjects were required to focus on repeating the keywords as accurately as possible, and we encouraged them to guess the uncertain words. Our observation was that most participants listened to each word two or three times before moving on. Vowel and consonant recognition were measured using a 16-alternative identification paradigm. The response buttons were labeled using vowel syllables for the vowel recognition task, consonant context with common finals for the consonant recognition task. Lexical tone recognition was measured using a four-alternative identification paradigm, and "Lexical tone 1," "Lexical tone 2," "Lexical tone 3," and "Lexical tone 4" for the Mandarin lexical tone recognition task. No feedback was given during the formal test. Each word was rated as correct or incorrect, and then the percentage of correct words was recorded under different conditions. Subjects can take a rest at any time to minimize fatigue during testing. The complete test time for each participant is ∼1.5–2 h.

Least-Squares Approach

To evaluate the weight of the five frequency regions in Mandarin phoneme and lexical tone recognition using acoustic temporal E cues, we calculated the weight of each frequency region using the least-squares approach previously used in other research (Kasturi et al., 2002). The strength of each frequency region was defined as a binary value of 0 or 1, depending on whether the frequency region was presented or not. Then, the weight of each frequency region was calculated by predicting the subject's response as a linear combination of the contribution of each frequency region. The initial weights of each subject's five frequency regions were converted to relative weights by summing them up, and the weights of each frequency region were expressed as the initial weight divided by the sum of all five frequency regions weights. Therefore, the weights of the five frequency regions add up to 1.0 (For more details, please see **Supplementary Material**).

Statistical Analysis

The Statistical Package for Social Sciences (SPSS) version 24.0 (IBM Corp., Armonk, NY, United States) was used for statistical analysis. The one-way analysis of variance (ANOVA) with repeated measures was used for the results from different test conditions for phoneme and lexical tone recognition. The *post hoc* analysis (Tukey's test) was used for pairwise comparison. The least-squares approach was used to calculate the relative weights of the five frequency regions. The independent samples *t*-test was used to compare the relative weights of five frequency regions in Mandarin phoneme and lexical tone recognition. The figures were generated by GraphPad Prism 8.0 (GraphPad Software, San Diego, CA, United States). Statistical significance was set at $p < 0.05$.

RESULTS

Scores for Mandarin Phoneme and Lexical Tone Recognition Across Conditions Using Temporal E Cues

As shown in **Figure 1A**, the vowel recognition scores ranged from 50.43 to 84.82% when the E cues were presented in three frequency regions. The Region 234 condition score was the highest, ∼84.82%, while Region 135 was lowest, ∼50.43%. A one-way repeated-measures ANOVA of different test conditions with three frequency regions showed significant differences in vowel recognition scores among different frequency regions combinations [$F_{(9,100)} = 13.559$, $p < 0.05$]. The Tukey's test revealed that the score under the Region 135 and Region 145 conditions was significantly lower than the scores under all other conditions with three frequency regions ($p < 0.05$). The consonant recognition scores ranged from 35.49 to 63.77% when the E cues were presented in three frequency regions (see in **Figure 1B**). The Region 345 condition score was the highest, ∼63.77%, while Region 123 was lowest, ∼35.49%. A one-way repeated-measures ANOVA of different test conditions with three frequency regions showed significant differences in consonant recognition scores among different frequency region combinations [$F_{(9,100)} = 11.622$, $p < 0.05$]. The Tukey's test revealed that the scores obtained from conditions combined with Frequency Region 5 would be higher than those obtained from conditions combined without Region 5 (Region 123, Region 124, Region 134, and Region 234) ($p < 0.05$). The lexical tone recognition scores ranged from 60.80 to 97.15% when the E cues were presented in three frequency regions (see in **Figure 1C**). The Region 124 condition score was the highest, ∼97.15%, while Region 345 was lowest, ∼60.80%. A one-way repeated-measures ANOVA of different test conditions with three frequency regions showed significant differences in lexical tone recognition scores among different frequency region combinations [$F_{(9,100)} = 46.910$, $p < 0.05$]. The Tukey's test revealed that the scores obtained from conditions combined with Frequency Region 1 would be higher than those obtained from conditions combined without Region 1 (Region 234, Region 235, Region 245, and Region 345) ($p < 0.05$).

As shown in **Figure 2A**, the vowel recognition scores ranged from 76.27 to 96.58% when the E cues were presented in four frequency regions. The Region 1234 condition score was the highest, ∼95.24%, while Region 1345 was the lowest, ∼76.27%. When stimulus presented in full frequency regions, the score raised to 96.58%. A one-way repeated-measures ANOVA of different test conditions with four and five frequency regions showed significant differences in vowel recognition scores among different frequency region combinations [$F_{(5,60)} = 27.674$, $p < 0.05$]. The Tukey's test revealed that the score under the Region 1345 condition was significantly lower than the score

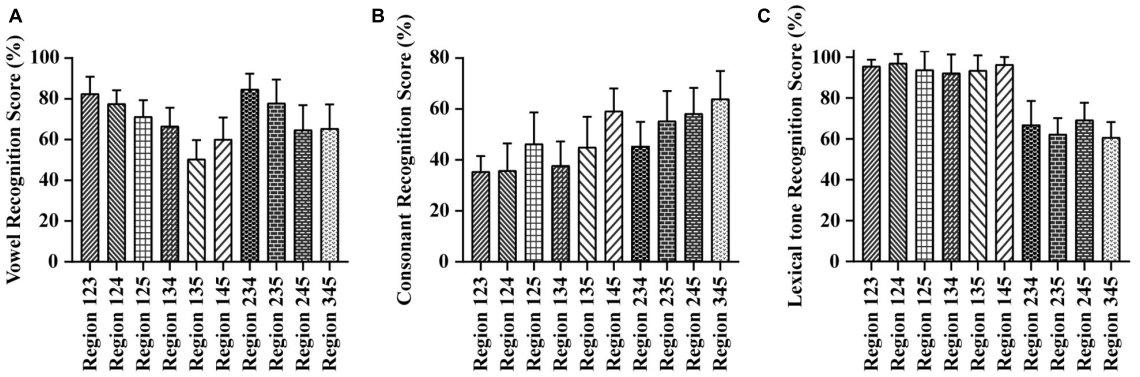

FIGURE 1 | Averaged percent-correct scores for Mandarin phoneme and lexical tone recognition using acoustic temporal envelope with three frequency regions conditions. The error bars represent standard errors. **(A)** Averaged scores for Mandarin vowel recognition with envelope cues in three frequency regions conditions. **(B)** Averaged scores for Mandarin consonant recognition with envelope cues in three frequency regions conditions. **(C)** Averaged scores for Mandarin lexical tone recognition with envelope cues in three frequency regions conditions.

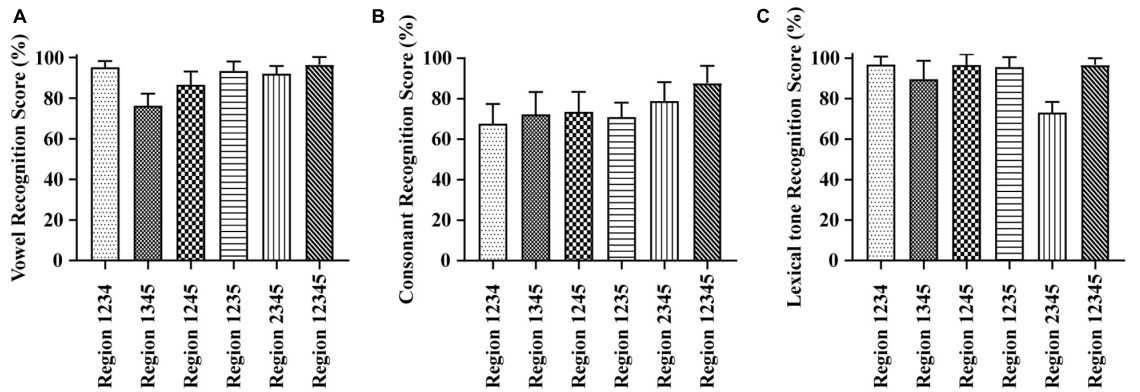

FIGURE 2 | Averaged percent-correct scores for Mandarin phoneme and lexical tone recognition using acoustic temporal envelope with four and five frequency regions conditions. The error bars represent standard errors. **(A)** Averaged scores for Mandarin vowel recognition with envelope cues in four frequency regions conditions. **(B)** Averaged scores for Mandarin consonant recognition with envelope cues in four frequency regions conditions. **(C)** Averaged scores for Mandarin lexical tone recognition with envelope cues in four frequency regions conditions.

under all other conditions ($p < 0.05$). The consonant recognition scores ranged from 67.75 to 78.87% when the E cues were presented in four frequency regions (see in **Figure 2B**). The Region 2345 condition score was the highest, ∼78.87%, while Region 1234 was the lowest, ∼67.75%. When stimulus presented in full frequency regions, the score raised to 87.87%. A one-way repeated-measures ANOVA of different test conditions with four and five frequency regions showed significant differences in consonant recognition scores among different frequency region combinations [$F_{(5,60)} = 6.462$, $p < 0.05$]. The Tukey's test revealed that the difference in the five conditions with four frequency regions was not significant ($p = 0.063$). However, the consonant recognition scores with full frequency regions were significantly higher than that in four frequency regions combinations ($p < 0.05$). The lexical tone recognition scores ranged from 73.16 to 96.87% when the E cues were presented in four frequency regions (see in **Figure 2C**). The Region 1234 condition score was the highest, ∼96.87%, while Region 2345 was lowest, ∼73.16%. When stimulus presented in full frequency regions, the score raised to 96.73%. A one-way repeated-measures ANOVA of different test conditions with four and five frequency regions showed significant differences in consonant recognition scores among different frequency region combinations [$F_{(5,60)} = 30.802$, $p < 0.05$]. The Tukey's test revealed that the score under the Region 2345 condition was significantly lower than the score under all other conditions with four frequency regions ($p < 0.05$).

Relative Weights of the Five Frequency Regions in Mandarin Phoneme and Lexical Tone Recognition

As shown in **Figure 3**, the mean weights of frequency region 1–5 for vowel recognition were 0.17, 0.31, 0.22, 0.18, and 0.12, respectively. The one-way ANOVA showed a significant main effect of region on weight for vowel recognition [$F_{(4,50)} = 41.117$, $p < 0.05$]. The Tukey's test revealed that the relative weight of Region 2 was highest than all other regions while the relative

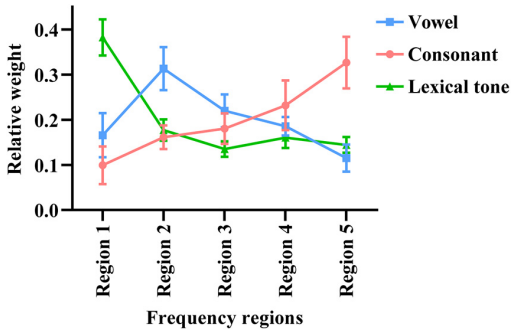

FIGURE 3 | The relative weights of different frequency regions for Mandarin phoneme and lexical tone recognition using acoustic temporal envelope. The error bars represent standard errors.

weight of Region 5 was lowest than all other regions ($p < 0.05$). The mean weights of frequency region 1–5 for consonant recognition were 0.10, 0.16, 0.18, 0.23, and 0.33, respectively. The one-way ANOVA showed a significant main effect of region on weight for consonant recognition [$F_{(4,50)} = 40.459$, $p < 0.05$]. The Tukey's test revealed that the relative weight of Region 5 was highest than all other regions while the relative weight of Region 1 was lowest than all other regions ($p < 0.05$). The mean weights of frequency region 1–5 for lexical tone recognition were 0.38, 0.18, 0.14, 0.16, and 0.14, respectively. The one-way ANOVA showed a significant main effect of region on weight for lexical tone recognition [$F_{(4,50)} = 176.725$, $p < 0.05$]. The Tukey's test revealed that the relative weight of Region 1 was highest than all other regions ($p < 0.05$).

DISCUSSION

This study was designed to explore the relative importance of acoustic E cues across different frequency regions for Mandarin phoneme and lexical tone recognition. Then we calculated the weight of each frequency region in Mandarin phoneme and lexical tone recognition by the least-squares approach as shown in **Figure 3**. Region 2 (502–1,022 Hz), Region 5 (3,856–7,562 Hz), and Region 1 (80–502 Hz) significantly contributed to Mandarin vowel, consonant, and lexical tone recognition, respectively.

Previous reports suggested that Mandarin phonemes and sentence recognition improved dramatically when the number of frequency regions increased from one to four (Fu et al., 1998), which was in line with results found in English speech recognition (Shannon et al., 1995). Our results affirm these findings, and that when presented in full region, the temporal E cues are sufficient to code for the recognition of Mandarin phoneme and lexical tone.

Vowels are important to the power of speech that is characterized by open vocal tract with sustained vocalization, low-frequency energy, and long duration (Chen et al., 2013; Chen and Chan, 2016). Sounds typically have four or five formants when it passes through the vocal tract. The first two formants determine the quality of vowels, while the last three formants determine the individual's unique timbre and influence the individual's vocal characteristics. When vowels are pronounced, the height of the tongue position corresponds to the first formant (F1), and the front and back of the tongue position correspond to the second formant (F2) (Lopes et al., 2018). Formant frequencies F1 and F2 have long been known to be crucial for encoding the phonetic identity of vowels (Carney et al., 2015; Fogerty, 2015). Hillenbrand et al. (1995) analyzed the data obtained from 45 men, 48 women, and 46 children and revealed that F1 (342–1,022 Hz) and F2 (910–3,081 Hz) were sufficient for vowel classification. It seems that the locations of the formants are dispersed optimally in the F1–F2 space, as described in dispersion theory (Schwartz et al., 1997). With the increase of the size of vowel systems, this dispersion leads to the consistencies among linguistic vowel systems in the appearance of vowel contrasts. While a study (Parikh and Loizou, 2005) reported that vowel recognition in noise is supported mainly by information about F1 along with some information about F2, another study (Xu et al., 2005) analyzed the information transmitted for acoustic features of vowels and found that the duration and F1 cues rather than F2 cues contributed substantially to vowel recognition. This is closer to a report (Traunmüller, 1981) that suggested the simultaneous distance between F1 and the fundamental frequency (F0) is the primary determinant of perceived vowel height. In our research, we found that the scores under the conditions without Region 2 (Region 135, Region 145, and Region 1345) were significantly lower than the score in other conditions (seen in **Figures 1A**, **2A**). The mean weights of frequency region 1–5 for vowel recognition were 0.17, 0.31, 0.22, 0.18, and 0.12, respectively. The relative weight of Region 2 (502–1,022 Hz) was highest across all other regions (seen in **Figure 3**). This is consistent with previously reported study (Kasturi et al., 2002) that channels 1, 3, and 4, centered at 393, 1,037, and 1,685 Hz, respectively, received the largest weight for vowels recognition and will lead to decrease in listener's performance if removed.

Different from vowels, consonants are characterized by complete or partial vocal tract constriction with high-frequency energy and short duration that are important to speech intelligibility (Chen et al., 2013; Chen and Chan, 2016). For consonants, many of these phonemes are characterized by rapid, instantaneous changes in amplitude, for instance, those caused by burst noise (Stevens, 2002). Therefore, high frequencies phoneme level modulation may be particularly important for conveying the consonant cues necessary for intelligibility. These high-frequency bands are characterized by having fast rate E modulations. A previous research (Fogerty, 2014) replaced consonant and vowel segments with noise matched to speech spectrum and found that consonants contain higher frequency components compared to vowels. We found that high-frequency region (3,856–7,562 Hz) of E cues plays a crucial role in consonant recognition (seen in **Figure 3**), and the scores obtained from conditions combined with Region 5 would be higher than those obtained from conditions combined without Region 5 (seen in **Figure 1B**). However, our result of relative weight for consonant recognition is different from previous findings (Kasturi et al., 2002). Here, the relative weight for the consonants was quite flat and all channels (the frequencies ranged from 300 to 4,444 Hz) are equally important for consonant recognition. In contrast, a study (Apoux and Bacon, 2008) reported that consonant recognition was not affected by removing E cues above 4 Hz

in the low- and high-frequency bands, while the consonant recognition decreased as the cutoff frequency was decreased in the mid-frequency region from 16 to 4 Hz. Possible reasons for the different weight for consonant recognition include: (1) differences in speech materials, for the type of speech material may have a strong impact on the value of acoustic information (Lunner et al., 2012); (2) the different methods of processing stimuli and different cut-off frequencies used in different experiments; (3) finally and most importantly the difference may come from differences in languages. The plosive bursts in Mandarin consonants produce a strongly synchronized burst of energy across the frequency spectrum related to high frequency regions (Drullman et al., 1994).

Lexical tone is also called pitch or the height of the sound. Most ordinary sounds can be analyzed as a sum of sinusoidal components with harmonic frequencies and evoke a pitch corresponding to their F0 (Santurette and Dau, 2011). As previous studies reported, F0 cues are important for Chinese lexical tone recognition (Luo and Fu, 2004; Chen et al., 2014; Vandali et al., 2015). There are four lexical tones in Mandarin including Lexical tone 1- (high-level), Lexical tone 2/(rising), Lexical tone 3 v (falling-rising), and Lexical tone 4 \(falling). As a tonal language, the same phonetic segment carries a different meaning when produced with different lexical tones (Chen et al., 2013). Lexical tones of Mandarin have been studied by many researchers. As previous research reported, Lexical tone 1 is associated with a flat F0 contour and Lexical tone 2 with a rising F0 contour (Whalen and Xu, 1992), Lexical tone 3 has the lowest intensity and longest duration while Lexical tone 4 is usually the strongest and shortest lasting pitch (Kuo et al., 2008). Although these four lexical tones are mainly distinguished by the F0 cues, other characteristics including overall intensity and duration vary systematically with lexical tone (Kuo et al., 2008). A study (Chen et al., 2014) examined the effects of lexical tone on the intelligibility of Mandarin revealed that the F0 contour is particularly important in tonal recognition in noisy environments. Another study (Vandali et al., 2015) reported that training with a single cue (F0 and center frequency) can improve the recognition ability of pitch and timbre without other cues variations. This is similar to another finding (Luo and Fu, 2004) that found modifying the amplitude E to make it closer to the F0 contour may be an effective method to improve the lexical tone recognition of Chinese CI wearers. Our findings are consistent with these previous studies, where Region 1 (80–502 Hz) significantly contributes to Mandarin lexical tone recognition (seen in **Figure 3**).

There are some limitations in our study. First, the age of the participants ranged from 21 to 27, so the result has limited explanation for other age groups including the infant and the elderly. Secondly, the subjects in this study received higher education. Larger-scale studies are needed to verify the results, and for further research. Thirdly, the preliminary results our study currently obtained will be used to guide more in-depth research, however, how signal processing or stimulation strategies for future CI systems will be influenced by current results is unclear. Perhaps CI users, in addition to normal hearing listeners, could further be recruited to increase the impact of this research.

CONCLUSION

(1) For Mandarin vowel recognition, Region 2 (502–1,022 Hz) which contained the first formant (*F*1) information contributed more than other regions.
(2) For Mandarin consonant recognition, Region 5 (3,856–7,562 Hz) contributed more than other regions.
(3) For Mandarin lexical tone recognition, Region 1 (80–502 Hz) which contained fundamental frequency (F0) information contributed more than other regions.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Shanghai Jiaotong University Affiliated Sixth People's Hospital Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ZZhe, KL, ZZha, DQ, and YF: conceptualization. KL, YG, and SH: methodology. KL, GF, and YG: data curation. YL, LX, CL, and SH: investigation. ZZhe: writing—original draft preparation. ZZha, DQ, and YF: writing—review and editing. YF: funding acquisition. ZZhe and YF: resources. ZZha: supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the National Natural Science Foundation of China (No. 81771015) and the Shanghai Municipal Commission of Science and Technology (Grant No. 18DZ2260200) and the International Cooperation and Exchange of the National Natural Science Foundation of China (Grant No. 81720108010).

ACKNOWLEDGMENTS

We would like to thank Qian-Jie Fu for software support, and we also would like to thank the participants of the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.744959/full#supplementary-material

REFERENCES

Apoux, F., and Bacon, S. P. (2004). Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise. *J. Acoust. Soc. Am.* 116, 1671–1680. doi: 10.1121/1.1781329

Apoux, F., and Bacon, S. P. (2008). Differential contribution of envelope fluctuations across frequency to consonant identification in quiet. *J. Acoust. Soc. Am.* 123, 2792. doi: 10.1121/1.2897916

Ardoint, M., Agus, T., Sheft, S., and Lorenzi, C. (2011). Importance of temporal-envelope speech cues in different spectral regions. *J. Acoust. Soc. Am.* 130, El115–El121. doi: 10.1121/1.3602462

Ardoint, M., and Lorenzi, C. (2010). Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues. *Hear. Res.* 260, 89–95. doi: 10.1016/j.heares.2009.12.002

Boëx, C., Pelizzone, M., and Montandon, P. (1996). Speech recognition with a CIS strategy for the ineraid multichannel cochlear implant. *Am. J. Otol.* 17, 61–68.

Buechner, A., Frohne-Buechner, C., Boyle, P., Battmer, R. D., and Lenarz, T. (2009). A high rate n-of-m speech processing strategy for the first generation Clarion cochlear implant. *Intern. J. Audiol.* 48, 868–875. doi: 10.3109/14992020903095783

Carney, L. H., Li, T., and McDonough, J. M. (2015). Speech coding in the brain: representation of vowel formants by midbrain meurons tuned to sound fluctuations. *eNeuro* 2:ENEURO.0004-15.2015. doi: 10.1523/eneuro.0004-15.2015

Chen, F., and Chan, F. W. (2016). Understanding frequency-compressed Mandarin sentences: role of vowels. *J. Acoust. Soc. Am.* 139, 1204–1213. doi: 10.1121/1.4944037

Chen, F., Wong, L. L., and Hu, Y. (2014). Effects of lexical tone contour on Mandarin sentence intelligibility. *J. Speech Lang. Hear. Res.* 57, 338–345. doi: 10.1044/1092-4388(2013/12-0324)

Chen, F., Wong, L. L., and Wong, E. Y. (2013). Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility. *J. Acoust. Soc. Am.* 134, El178–El184. doi: 10.1121/1.4812820

Drullman, R., Festen, J. M., and Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.* 95, 1053–1064. doi: 10.1121/1.408467

Fogerty, D. (2014). Importance of envelope modulations during consonants and vowels in segmentally interrupted sentences. *J. Acoust. Soc. Am.* 135, 1568–1576. doi: 10.1121/1.4863652

Fogerty, D. (2015). Indexical properties influence time-varying amplitude and fundamental frequency contributions of vowels to sentence intelligibility. *J. Phonet.* 52, 89–104. doi: 10.1016/j.wocn.2015.06.005

Fu, Q. J., Zeng, F. G., Shannon, R. V., and Soli, S. D. (1998). Importance of tonal envelope cues in Chinese speech recognition. *J. Acoust. Soc. Am.* 104, 505–510. doi: 10.1121/1.423251

Glasberg, B. R., and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* 47, 103–138. doi: 10.1016/0378-5955(90)90170-t

Guo, Y., Sun, Y., Feng, Y., Zhang, Y., and Yin, S. (2017). The relative weight of temporal envelope cues in different frequency regions for Mandarin sentence recognition. *Neural Plast.* 2017:7416727. doi: 10.1155/2017/7416727

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97(5 Pt 1), 3099–3111. doi: 10.1121/1.411872

Jing, Y., Yu, Z., Li, A., and Li, X. (2017). "On the duration of mandarin tones," in *Proceedings of the Interspeech 2017*, Stockholm.

Kasturi, K., Loizou, P. C., Dorman, M., and Spahr, T. (2002). The intelligibility of speech with "holes" in the spectrum. *J. Acoust. Soc. Am.* 112(3 Pt 1), 1102–1111. doi: 10.1121/1.1498855

Kewley-Port, D., Burkle, T. Z., and Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *J. Acoust. Soc. Am.* 122, 2365–2375. doi: 10.1121/1.2773986

Kong, Y. Y., and Zeng, F. G. (2006). Temporal and spectral cues in Mandarin tone recognition. *J. Acoust. Soc. Am.* 120(5 Pt 1), 2830–2840. doi: 10.1121/1.2346009

Kuo, Y. C., Rosen, S., and Faulkner, A. (2008). Acoustic cues to tonal contrasts in Mandarin: implications for cochlear implants. *J. Acoust. Soc. Am.* 123:2815. doi: 10.1121/1.2896755

Li, B., Hou, L., Xu, L., Wang, H., Yang, G., Yin, S., et al. (2015). Effects of steep high-frequency hearing loss on speech recognition using temporal fine structure in low-frequency region. *Hear. Res.* 326, 66–74. doi: 10.1016/j.heares.2015.04.004

Li, B., Wang, H., Yang, G., Hou, L., Su, K., Feng, Y., et al. (2016). The importance of acoustic temporal fine structure cues in different spectral regions for Mandarin sentence recognition. *Ear Hear.* 37, e52–e56. doi: 10.1097/aud.0000000000000216

Lopes, L. W., Alves, J. D. N., Evangelista, D. D. S., França, F. P., Vieira, V. J. D., Lima-Silva, M. F. B., et al. (2018). Accuracy of traditional and formant acoustic measurements in the evaluation of vocal quality. *Codas* 30:e20170282. doi: 10.1590/2317-1782/20182017282

Lunner, T., Hietkamp, R. K., Andersen, M. R., Hopkins, K., and Moore, B. C. (2012). Effect of speech material on the benefit of temporal fine structure information in speech for young normal-hearing and older hearing-impaired participants. *Ear Hear.* 33, 377–388. doi: 10.1097/AUD.0b013e3182387a8c

Luo, X., and Fu, Q. J. (2004). Enhancing Chinese tone recognition by manipulating amplitude envelope: implications for cochlear implants. *J. Acoust. Soc. Am.* 116, 3659–3667. doi: 10.1121/1.1783352

Macherey, O., and Carlyon, R. P. (2014). Cochlear implants. *Curr. Biol.* 24, R878–R884. doi: 10.1016/j.cub.2014.06.053

Meng, Q., Zheng, N., and 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, 301–310. doi: 10.1121/1.4939707

Nissen, S. L., Harris, R. W., Jennings, L. J., Eggett, D. L., and Buck, H. (2005). Psychometrically equivalent Mandarin bisyllabic speech discrimination materials spoken by male and female talkers. *Intern. J. Audiol.* 44, 379–390. doi: 10.1080/14992020500147615

Parikh, G., and Loizou, P. C. (2005). The influence of noise on vowel and consonant cues. *J. Acoust. Soc. Am.* 118, 3874–3888. doi: 10.1121/1.2118407

Pfingst, B. E., Bowling, S. A., Colesa, D. J., Garadat, S. N., Raphael, Y., Shibata, S. B., et al. (2011). Cochlear infrastructure for electrical hearing. *Hear. Res.* 281, 65–73. doi: 10.1016/j.heares.2011.05.002

Psarros, C. E., Plant, K. L., Lee, K., Decker, J. A., Whitford, L. A., and Cowan, R. S. (2002). Conversion from the SPEAK to the ACE strategy in children using the nucleus 24 cochlear implant system: speech perception and speech production outcomes. *Ear Hear.* 23(Suppl. 1), 18S–27S. doi: 10.1097/00003446-200202001-00003

Qi, B., Mao, Y., Liu, J., Liu, B., and Xu, L. (2017). Relative contributions of acoustic temporal fine structure and envelope cues for lexical tone perception in noise. *J. Acoust. Soc. Am.* 141:3022. doi: 10.1121/1.4982247

Roy, A. T., Carver, C., Jiradejvong, P., and Limb, C. J. (2015). Musical sound quality in cochlear implant users: a comparison in bass frequency perception between fine structure processing and high-definition continuous interleaved sampling strategies. *Ear Hear.* 36, 582–590. doi: 10.1097/aud.0000000000000170

Santurette, S., and Dau, T. (2011). The role of temporal fine structure information for the low pitch of high-frequency complex tones. *J. Acoust. Soc. Am.* 129, 282–292. doi: 10.1121/1.3518718

Schwartz, J.-L., Boë, L.-J., Vallée, N., and Abry, C. (1997). The dispersion-focalization theory of vowel systems. *J. Phonet.* 25, 255–286. doi: 10.1006/jpho.1997.0043

Shannon, R. V., Galvin, J. J. III, and Baskent, D. (2002). Holes in hearing. *J. Assoc. Res. Otolaryngol.* 3, 185–199. doi: 10.1007/s101620020021

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science* 270, 303–304. doi: 10.1126/science.270.5234.303

Skinner, M. W., Holden, L. K., Whitford, L. A., Plant, K. L., Psarros, C., and Holden, T. A. (2002). Speech recognition with the nucleus 24 SPEAK, ACE, and CIS speech coding strategies in newly implanted adults. *Ear Hear.* 23, 207–223. doi: 10.1097/00003446-200206000-00005

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature* 416, 87–90. doi: 10.1038/416087a

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.* 111, 1872–1891. doi: 10.1121/1.1458026

Tabibi, S., Kegel, A., Lai, W. K., and Dillier, N. (2020). A bio-inspired coding (BIC) strategy for cochlear implants. *Hear. Res.* 388:107885. doi: 10.1016/j.heares.2020.107885

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *J. Acoust. Soc. Am.* 69, 1465–1475. doi: 10.1121/1.385780

Vandali, A., Dawson, P., Au, A., Yu, Y., Brown, M., Goorevich, M., et al. (2019). Evaluation of the optimized pitch and language strategy in cochlear implant recipients. *Ear Hear.* 40, 555–567. doi: 10.1097/aud.0000000000000627

Vandali, A., Sly, D., Cowan, R., and van Hoesel, R. (2015). Training of cochlear implant users to improve pitch perception in the presence of competing place cues. *Ear Hear.* 36, e1–e13. doi: 10.1097/aud.0000000000000109

Wang, S., Xu, L., and Mannell, R. (2011). Relative contributions of temporal envelope and fine structure cues to lexical tone recognition in hearing-impaired listeners. *J. Assoc. Res. Otolaryngol.* 12, 783–794. doi: 10.1007/s10162-011-0285-0

Wang, W., Zhou, N., and Xu, L. (2011). Musical pitch and lexical tone perception with cochlear implants. *Intern. J. Audiol.* 50, 270–278. doi: 10.3109/14992027.2010.542490

Warren, R. M., Bashford, J. A. Jr., and Lenz, P. W. (2004). Intelligibility of bandpass filtered speech: steepness of slopes required to eliminate transition band contributions. *J. Acoust. Soc. Am.* 115, 1292–1295. doi: 10.1121/1.1646404

Wei, C. G., Cao, K., and Zeng, F. G. (2004). Mandarin tone recognition in cochlear-implant subjects. *Hear. Res.* 197, 87–95. doi: 10.1016/j.heares.2004.06.002

Whalen, D. H., and Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica* 49, 25–47. doi: 10.1159/000261901

Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., and Rabinowitz, W. M. (1991). Better speech recognition with cochlear implants. *Nature* 352, 236–238. doi: 10.1038/352236a0

World Health Organization [WHO] (2020). Available online at: https://www.who.int/news-room/factsheets/detail/deafness-and-hearing-loss (accessed April 1, 2021).

Wu, J. L., Yang, H. M., Lin, Y. H., and Fu, Q. J. (2007). Effects of computer-assisted speech training on Mandarin-speaking hearing-impaired children. *Audiol. Neurootol.* 12, 307–312. doi: 10.1159/000103211

Xu, L., and Pfingst, B. E. (2003). Relative importance of temporal envelope and fine structure in lexical-tone perception. *J. Acoust. Soc. Am.* 114(6 Pt 1), 3024–3027. doi: 10.1121/1.1623786

Xu, L., Thompson, C. S., and Pfingst, B. E. (2005). Relative contributions of spectral and temporal cues for phoneme recognition. *J. Acoust. Soc. Am.* 117, 3255–3267. doi: 10.1121/1.1886405

Zeng, F. G. (2004). Trends in cochlear implants. *Trends Amplif.* 8, 1–34. doi: 10.1177/108471380400800102

Zeng, F. G., Rebscher, S., Harrison, W., Sun, X., and Feng, H. (2008). Cochlear implants: system design, integration, and evaluation. *IEEE Rev. Biomed. Eng.* 1, 115–142. doi: 10.1109/rbme.2008.2008250

Zheng, Z., Li, K., Guo, Y., Wang, X., Xiao, L., Liu, C., et al. (2021). The relative weight of temporal envelope cues in different frequency regions for Mandarin disyllabic word recognition. *Front. Neurosci.* 15:670192. doi: 10.3389/fnins.2021.670192

Ziese, M., Stützel, A., von Specht, H., Begall, K., Freigang, B., Sroka, S., et al. (2000). Speech understanding with the CIS and the n-of-m strategy in the MED-EL COMBI 40+ system. *J. Otorhinolaryngol. Relat. Spec.* 62, 321–329. doi: 10.1159/000027763

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

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