frontiers in Neuroscience

ORIGINAL RESEARCH
published: 15 July 2021
doi: 10.3389/fnins.2021.670192

The Relative Weight of Temporal Envelope Cues in Different Frequency Regions for Mandarin Disyllabic Word Recognition

Zhong Zheng[1,2], Keyi Li[3], Yang Guo[4], Xinrong Wang[1,2], Lili Xiao[1,2], Chengqi Liu[1,2], Shouhuan He[5], Gang Feng[6] 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] Ear, Nose, and Throat Institute and Otorhinolaryngology Department, Eye and ENT Hospital of Fudan University, Shanghai, China, [5] Department of Otolaryngology, Qingpu Branch of Zhongshan Hospital Affiliated to Fudan University, Shanghai, China, [6] The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, China

OPEN ACCESS

Edited by:
Stephen Charles Van Hedger, Huron University College, Canada

Reviewed by:
Christian Füllgrabe, Loughborough University, United Kingdom
Nan Yan, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences (CAS), China

***Correspondence:**
Gang Feng
fghyc@163.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: 20 February 2021
Accepted: 14 June 2021
Published: 15 July 2021

Citation:
Zheng Z, Li K, Guo Y, Wang X, Xiao L, Liu C, He S, Feng G and Feng Y (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

Objectives: Acoustic temporal envelope (E) cues containing speech information are distributed across all frequency spectra. To provide a theoretical basis for the signal coding of hearing devices, we examined the relative weight of E cues in different frequency regions for Mandarin disyllabic word recognition in quiet.

Design: E cues were extracted from 30 continuous frequency bands within the range of 80 to 7,562 Hz using Hilbert decomposition and assigned to five frequency regions from low to high. Disyllabic word recognition of 20 normal-hearing participants were obtained using the E cues available in two, three, or four frequency regions. The relative weights of the five frequency regions were calculated using least-squares approach.

Results: Participants correctly identified 3.13–38.13%, 27.50–83.13%, or 75.00–93.13% of words when presented with two, three, or four frequency regions, respectively. Increasing the number of frequency region combinations improved recognition scores and decreased the magnitude of the differences in scores between combinations. This suggested a synergistic effect among E cues from different frequency regions. The mean weights of E cues of frequency regions 1–5 were 0.31, 0.19, 0.26, 0.22, and 0.02, respectively.

Conclusion: For Mandarin disyllabic words, E cues of frequency regions 1 (80–502 Hz) and 3 (1,022–1,913 Hz) contributed more to word recognition than other regions, while frequency region 5 (3,856–7,562) contributed little.

Keywords: relative weight, envelope cues, frequency region, Mandarin Chinese, disyllabic word

INTRODUCTION

The World Health Organization estimates that > 5% of the world's population (approximately 466 million people) suffer from disabling hearing loss (WHO, 2020). Approximately one-third of people over the age of 65 years suffer from different degrees of sensorineural hearing loss (SNHL), one of the most common forms of hearing loss (WHO, 2020). Cochlear implants (CIs) remain the only

effective device for restoring speech communication ability in patients with severe to profound SNHL (Tavakoli et al., 2015). In people with normal hearing, the cochlea converts speech signals into bioelectrical signals, which are transmitted through the auditory nerve to the brain so that listeners are able to sense various sounds. The basilar membrane in the cochlea can be regarded as a series of overlapping bandpass filters, each of which has its own unique characteristic frequency. When the basilar membrane is vibrated by sound, the sound signal of a characteristic frequency causes amplitude to peak at the corresponding basilar membrane partition (Smith et al., 2002). As electronic devices, CIs stimulate the auditory nerve *via* amplitude-modulated pulses that carry important temporal information of speech signals.

Speech is a complex acoustic signal and can be viewed in terms of two domains: temporal information and spectral information. Temporal information refers to information in speech signals with time-varying wave rates, which can be divided into the temporal envelope (E) below 50 Hz, periodic fluctuations in the range of 50–500 Hz, and temporal fine structure in the range of 500–10,000 Hz (Rosen, 1992). E cues contain temporal modulation information, which is most important for speech perception in quiet conditions, whereas the temporal fine structure can provide information in noisy environments and for tonal and pitch recognition (Smith et al., 2002; Xu and Pfingst, 2003; Moore, 2008; Ardoint and Lorenzi, 2010; Wang et al., 2016). Vocoder studies have shown that E modulation rates of 4–16 Hz are most important for speech intelligibility in quiet (Drullman et al., 1994a,b; Shannon et al., 1995). However, when speech has to be perceived against interfering speech, both slower and faster modulation rates, which are associated with prosodic (Füllgrabe et al., 2009) and fundamental frequency (Xu et al., 2002; Stone et al., 2008), respectively, become important for identification.

Recently, many researchers have investigated the relative importance of E cues from different frequency regions. Shannon et al. (2002) studied the influence of temporal E cues from different frequency regions on English recognition by removing specific spectral information. They found that the removal medium and high frequencies had greater impacts than low frequencies. This is supported by Apoux and Bacon (2004) who also reported that temporal E cues from different frequency regions are important in quiet conditions, while the high-frequency region (>2,500 Hz) is more important in noisy environments. In addition, using a classic high-pass and low-pass filtering experiment paradigm (French and Steinberg, 2005), Ardoint et al. (Ardoint and Lorenzi, 2010) demonstrated that E cues for frequency bands approximately 1,000–2,000 Hz are most important in French vowel–consonant–vowel speech recognition. The different frequency ranges and their respective importance in these investigations of non-tonal languages motivated us to examine the relative weight of E cues in different frequency regions for Mandarin Chinese speech recognition.

Mandarin Chinese is a tonal language and has the most first-language speakers of any language in the world. It includes 23 consonants, 38 vowels, and 5 tones. The 38 vowels consist of 9 monophthongs, 13 diphthongs and triphthongs, and 16 nasal finals. The tones include tone 1 (high-level), tone 2 (mid-rising), tone 3 (low-dipping), and tone 4 (high-falling) (Xu and Zhou, 2011). In addition, there is a fifth tone, usually called a neutral tone or tone 0, that occurs in unstressed syllables in multisyllabic words or connected speech (Yang et al., 2017). There are many polysyllabic words in Mandarin, most of which are disyllabic, and different tones can represent many different meanings (Nissen et al., 2005). Thus, the recognition of disyllabic words plays an important role in Mandarin speech recognition. Despite the large number of people who speak Mandarin as a mother tongue, there has been little emphasis on the relative weight of temporal and spectral information in different frequency regions for Mandarin Chinese. The present study intends to fill this knowledge gap. Our results may benefit CI wearers whose native language is Mandarin Chinese and ultimately improve their speech recognition performance and quality of life (McRackan et al., 2018).

Speech recognition is an interactive process between the speech characteristics of auditory signals and the long-term language knowledge of the listener, enabled by the decoding of speech through integrative bottom-up and top-down processes (Tuennerhoff and Noppeney, 2016). Speech recognition includes phonemes, syntax, and semantic recognition (Etchepareborda, 2003; Desroches et al., 2009). Phonemes are the smallest unit of sound that distinguish one word from another word in a language. Syntax refers to the meaning and interpretation of words, signs, and sentence structure, and depends on elements such as language environment and contextual information. Based on the semantic information of language, we can roughly judge the content range. Bottom-up mechanisms include phoneme recognition (Tuennerhoff and Noppeney, 2016), which disyllabic word recognition primarily relies on. Top-down mechanisms include syntactic and semantic information (Etchepareborda, 2003; Desroches et al., 2009). In an fMRI study, Tuennerhoff and Noppeney (2016) found that activations of unintelligible fine-structure speech were limited to the primary auditory cortices, but when top-down mechanisms made speech intelligible, the activation spread to posterior middle temporal regions, allowing for lexical access and speech recognition. Sentences can provide listeners with an envelope template where lexical and phonological constraints can help segment the acoustic signal into larger comprehensible temporal units, similar to the spatial "pop-out" phenomenon in visual object recognition (Dolan et al., 1997). Both top-down and bottom-up mechanisms are essential in the recognition of sentences.

While Mandarin Chinese word recognition relies more on tone recognition, sentence recognition can be inferred from context, which is consistent with the top-down mechanisms of speech recognition. Our team previously found that acoustic temporal E cues in frequency regions 80–502 Hz and 1,022–1,913 Hz contributed significantly to Mandarin sentence recognition (Guo et al., 2017). The present study builds on these findings and further investigates the relative weights of E cues for Mandarin disyllabic word recognition.

MATERIALS AND METHODS

Participants

We recruited a total of 20 participants (10 males and 10 females), who were graduates of Shanghai Jiao Tong University with normal audiometric thresholds ($\leq$ 20 dB HL) bilaterally at octave frequencies of 0.25–8 kHz. Their ages ranged from 22 to 28 (average, 24) years. All participants were native Mandarin speakers with no reported history of ear disease or hearing loss. Pure-tone audiometric thresholds were measured with a GSI-61 audiometer (Grason-Stadler, Madison, WI, United States) using standard audiometric procedures. No participant had any preceding exposure to the speech materials used in the present study. All participants provided signed consent forms before the experiment and were compensated on an hourly basis for their participation. The protocol was approved by the ethics committee of Shanghai Jiao Tong University Affiliated Sixth People's Hospital (ChiCTR-ROC-17013460), and the experiment was performed in accordance with the Declaration of Helsinki.

Signal Processing

Mandarin disyllabic speech test materials issued by the Beijing Institute of Otolaryngology were used for disyllable word recognition in quiet conditions. The test materials included 10 lists, each of which contained 50 disyllabic words covering 96.65% of the words used in daily life (Wang et al., 2007). The disyllable words were filtered into 30 contiguous frequency bands using zero-phase, third-order Butterworth filters (18 dB/octave slopes), ranging from 80 to 7,562 Hz (Li et al., 2016). Each band had an equivalent rectangular bandwidth for normal-hearing participants, simulating the frequency selectivity of the normal auditory system (Glasberg and Moore, 1990).

The E cues were extracted from each band using Hilbert decomposition followed by low-pass filtering at 64 Hz with a third-order Butterworth filter. E cues were then used to modulate the amplitude of a white-noise carrier. The modulated noise was filtered using the same bandpass filters and summed across frequency bands to form the frequency regions of acoustic E cues. We focused on the parameters used in clinics to assess hearing levels; i.e., 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) regions. We chose cutoff frequencies closest to the audiometric frequencies 500, 1,000, 2,000, 4,000, and 8,000 Hz. Thus, frequency bands 1–8, 9–13, 14–18, 19–24, and 25–30 were combined to form frequency regions 1–5 (**Table 1**).

TABLE 1 | Cutoff frequencies of the 30 frequency bands.

Frequency regions	Band	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

To investigate the role of different frequency regions in Mandarin disyllabic word recognition, participants were presented with acoustic E cues from two frequency regions (10 conditions, namely, Region 1 + 2, Region 1 + 3, Region 1 + 4, Region 1 + 5, Region 2 + 3, Region 2 + 4, Region 2 + 5, Region 3 + 4, Region3 + 5, and Region 4 + 5), three frequency regions (10 conditions, namely, Region 1 + 2 + 3, Region 1 + 2 + 4, Region 1 + 2 + 5, Region 1 + 3 + 4, Region 1 + 3 + 5, Region 1 + 4 + 5, Region 2 + 3 + 4, Region 2 + 3 + 5, Region 2 + 4 + 5, and Region 3 + 4 + 5), and four frequency regions (5 conditions, namely, Region 1 + 2 + 3 + 4, Region 1 + 3 + 4 + 5, Region 1 + 2 + 4 + 5, Region 1 + 2 + 3 + 5, and Region 2 + 3 + 4 + 5). To prevent the possible use of transitional band information (Warren et al., 2004; Li et al., 2015), the frequency regions including disyllabic word E cues and complementary frequency regions (i.e., noise-masking), were combined to present a speech-to-noise ratio of 16 dB. For example, the "Region 1 + 2" condition indicates that the participant was presented with a stimulus consisting of E cues for disyllabic words in frequency regions 1 and 2 with noise in frequency regions 3, 4, and 5. Similarly, "Region 1 + 2 + 3 + 4" indicates that the stimulus comprised acoustic temporal E cues in frequency regions 1, 2, 3, and 4 with noise in frequency region 5.

Testing Procedure

None of the participants had previously participated in perceptual experiments testing acoustic temporal E cues. The participants were tested individually in a double-walled, soundproof room. All stimuli were delivered *via* HD 205 II circumaural headphones (Sennheiser, Wedemark, Germany) at approximately 65 dB SPL (range, 60–75 dB SPL).

About 30 min of practice was provided before the formal test. The practice vocabulary comprised 50 disyllabic words (i.e., one list of the test material). Words were first presented under the "full region" condition and then experimental stimuli were presented randomly. Feedback was given during practice. To familiarize participants with the stimuli, they were allowed to listen to words repeatedly for any number of times before moving on to the next word until their performance stabilized.

In the formal test, participants were allowed to listen to words as many times as they wanted. Most participants listened to each word two or three times before moving on. All conditions and corresponding material lists were presented in random order for each participant to avoid any order effect. Participants were encouraged to repeat words as accurately as possible and to guess if necessary. No feedback was provided during the test. Each keyword in the disyllabic word lists was rated as correct or incorrect, and the results were recorded as the percentage of correct words under different conditions. The participants were allowed to take breaks whenever necessary. Each participant required approximately 1.5–2 h to complete the set of tests.

Statistical Analysis

Statistical analyses were conducted using the Statistical Package for Social Sciences version 22.0 (IBM Co., Armonk, NY, United States). Because of our small sample size, we used the Kruskal–Wallis test to analyze results from different test conditions for disyllabic word recognition. Pairwise comparison using results of different test conditions for disyllabic word recognition was performed with *post hoc* analysis (the Bonferroni test). The relative weights of the five frequency regions were calculated using the least-squares approach (Kasturi et al., 2002). The Mann–Whitney *U* test was used to compare the relative weights of the five frequency regions for Mandarin disyllabic word and sentence recognition.

RESULTS

Scores for Mandarin Disyllabic Word Recognition Across Conditions Using E Cues

Participants correctly identified 3.13–38.13% of words when presented with E cues in two frequency regions (**Figure 1**); scores were highest for Region 1 + 2 (38.13% correct) and lowest for Region 2 + 4 (3.13% correct). Thus, these conditions were unfavorable for participants to understand the meaning of disyllabic words. In addition, the percentage of correct responses differed significantly among frequency region combinations (H = 167.288, $p < 0.05$). Regions 1 + 2 and Region 1 + 3 had significantly higher scores than other conditions with two frequency regions (adjusted $p < 0.05$).

Participants correctly identified 27.50%–83.13% of words when presented E cues in three frequency regions (**Figure 2**). Region 1 + 2 + 3 and Regions 2 + 3 + 4 had high scores (78.75% and 83.13% correct, respectively), while Region 2 + 3 + 5 scored relatively low (27.50% correct). In addition, the percentage of correct responses differed significantly among frequency region combinations (H = 168.938, $p < 0.05$). Region 1 + 2 + 3 and Region 2 + 3 + 4 had significantly higher scores than other conditions with three frequency regions (adjusted $p < 0.05$).

Participants correctly identified 75.00–93.13% of words when presented with E cues in four frequency regions (**Figure 3**). Region 1 + 2 + 3 + 4 had the highest score among all conditions (93.13% correct), while Region 1 + 2 + 4 + 5 had the lowest score among combinations of four frequency regions (75.00% correct). In addition, the percentage of correct responses differed

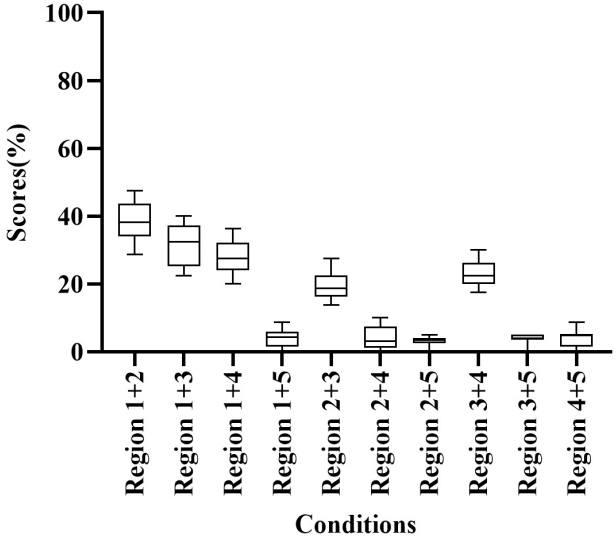

FIGURE 1 | Percent-correct scores for disyllabic word recognition using acoustic temporal envelope in two-frequency-region conditions.

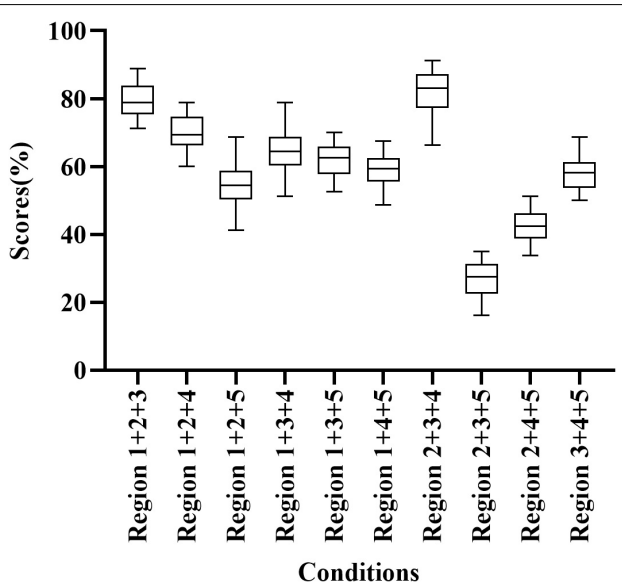

FIGURE 2 | Percent-correct scores for disyllabic word recognition using acoustic temporal envelope in three-frequency-region conditions.

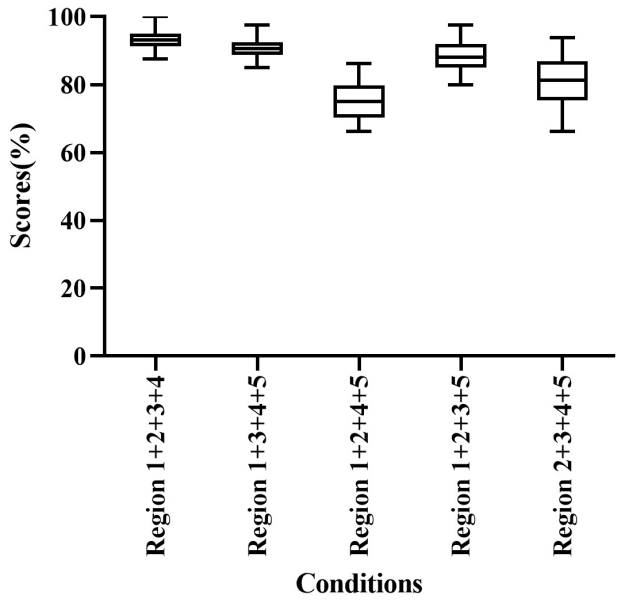

FIGURE 3 | Percent-correct scores for disyllabic word recognition using acoustic temporal envelope in four-frequency-region conditions.

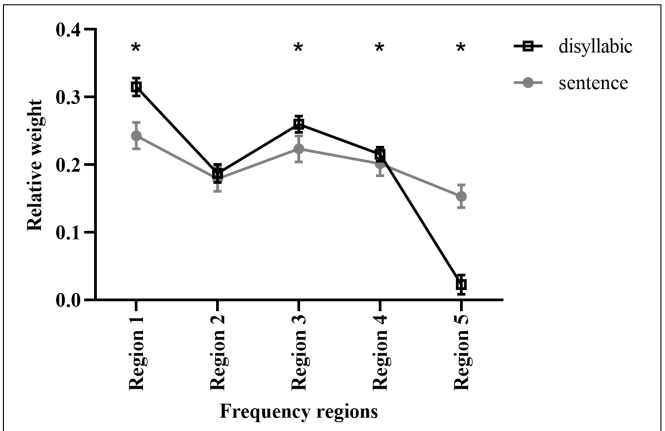

FIGURE 4 | The relative weights of different frequency regions for Mandarin disyllabic word and sentence recognition using acoustic temporal envelope. The data for Mandarin sentence recognition was adopted from a previous study (Guo et al., 2017). The error bars represent standard errors.
* Statistically significant ($p < 0.05$).

significantly among frequency region combinations (H = 60.762, $p < 0.05$). As the number of frequency regions increased, speech recognition scores increased and the magnitude of the difference between the groups decreased. Region 1 + 2 + 3 + 4 and Region 1 + 3 + 4 + 5 had significantly higher scores than other conditions with four frequency regions (adjusted $p < 0.05$).

Relative Weights of the Five Frequency Regions in Mandarin Disyllabic Word and Sentence Recognition

The relative weights of the five frequency regions for Mandarin disyllabic word recognition using acoustic temporal E cues were calculated using the least-squares approach (Kasturi et al., 2002). The strength of each frequency region was defined as a binary value (0 or 1) depending on whether the frequency region was present. The weight of each frequency region was then calculated by predicting the participant's response as a linear combination of each frequency region's intensity. The original weights for the five frequency regions of each participant were transformed to relative weights by summing their values, and each frequency region weight was represented as the original weight divided by the sum of all weights of the five frequency regions. Thus, the sum of the relative weights of the five frequency regions was equal to 1.0. The mean relative weights of frequency regions 1–5 were 0.31, 0.19, 0.26, 0.22, and 0.02, respectively (**Figure 4**). The relative weights differed significantly among frequency regions (H = 94.221, $p < 0.05$).

Our previous reports showed that the relative weights of the E cues from frequency regions 1–5 for Mandarin sentence recognition were 0.25, 0.18, 0.22, 0.20, and 0.15, respectively (Guo et al., 2017). Thus, the observed trends between sentence and disyllabic word recognition were similar. The E cues of frequency regions 1 and 3 contributed more to Mandarin disyllabic word and sentence recognition than those of the other frequency regions. Mandarin disyllabic word and sentence recognition had significantly different relative weights for frequency regions 1, 3, 4, and 5 ($p < 0.05$; **Figure 4**) but not for frequency region 2.

DISCUSSION

Vocoder technology is often used to simulate the signal processing of CIs when studying the function of E cues in speech recognition. The vocoder separates the speech signal into different frequency bands through bandpass filters. The E cues in different frequency bands are extracted and used to modulate white noise or sine waves. Eventually, the summed E cues of each frequency band are presented to the participants for perceptual experiments (Kim et al., 2015; Rosen and Hui, 2015). Previous research has shown that as the number of frequency bands segmented by a vocoder increases, the speech recognition rate of the subjects gradually increases (Shannon et al., 2004; Xu et al., 2005; Xu and Zheng, 2007). However, it is impossible to increase the number of intracochlear electrodes infinitely due to a number of constraints, including interference between adjacent electrodes (Shannon et al., 2004). Therefore, when the number of electrodes is fixed, it is necessary to study the relative importance of E cues in different frequency regions for speech recognition, especially given the trade-off between the number of spectral channels and the amount of temporal information (Xu and Pfingst, 2008).

It is worth noting that the bandpass filters used in the present study had a fairly shallow slope (18 dB/octave). The listening ability beyond the cutoff frequency that filtered the stimulus is called "off-frequency listening." However, our normal-hearing participants may not have been able to efficiently use the speech cues in the off-frequency bands because, in

vocoder processing, the fine structure of each frequency band is replaced by white noise.

The present study explored the relative importance of E cues across different frequency regions for Mandarin disyllabic word recognition. We found that E cues in different frequency regions contributed differentially to recognition scores. For Mandarin Chinese disyllabic words, frequency region 1 had the highest weight (0.31; **Figure 4**). Thus, the low-frequency region is most important for Mandarin recognition, which is consistent with a previous report on Mandarin Chinese sentence recognition (Guo et al., 2017). However, this is not consistent with the results of Ardoint et al., who found that E cues in higher regions (1,800–7,300 Hz) contributed more strongly to French consonant recognition (Ardoint et al., 2011). Explanations for this difference include differences in speech materials, since previous studies have shown that the type of speech material may have a strong impact on the use of acoustic temporal fine structure information (Lunner et al., 2012); the cutoff frequencies defined by different frequency regions used in different experiments; the methods of processing stimuli and evaluating weights; and, most importantly, differences inherent to the languages themselves. As a tone language, Mandarin Chinese has disyllabic words with different tones that can contain different lexical meanings (Fu et al., 1998; Wei et al., 2004). Fundamental frequency (F0) and its harmonics are the primary cues for lexical tone recognition, and tone contours play an important role in tonal language speech intelligibility (Feng et al., 2012). Kuo et al. (2008) found that when F0 information is provided, participants consistently have tone recognition rates of $> 90\%$; however, without F0 information, E cue information contributes to tone recognition to a lesser extent. Considering that F0 information is mainly in the low-frequency region, which plays an important role in tone recognition (Wong et al., 2007), the low-frequency region (i.e., region 1) likely has a strong influence on the recognition of Mandarin Chinese.

The E cues information in the intermediate frequencies (i.e., region 3, 1,022–1,913 Hz) is important for speech recognition, which is consistent with previous results. Kasturi et al. (2002) found that when E cues in a band centered at 1,685 Hz were removed, English vowel and consonant recognition declined. In addition, Ardoint et al. (Ardoint and Lorenzi, 2010) found that E cues conveyed important distinct phonetic cues in frequency regions between 1,000 and 2,000 Hz. These results indicate that E cues in the frequency band 1,000–2,000 Hz are important for speech recognition regardless of language.

We found that relative weights differed significantly between Mandarin disyllabic word and sentence recognition in frequency regions 1, 3, 4, and 5 ($p < 0.05$) but not in frequency region 2. This may have been due to differences in speech materials as described earlier and the fact that tone plays a more important role in understanding disyllabic words (Feng et al., 2012; Wong and Strange, 2017). Auditory speech input is rapidly and automatically bound by the participant into a speech representation in a short-term memory buffer. If this information matches the speech representation in long-term memory, relatively automatic and effortless lexical acquisition occurs. Mismatches are effortlessly controlled using higher-level linguistic knowledge such as semantics or syntactic contexts. Therefore, when bottom-up processes fail, controlled processing and sentence contextual information are used (Rönnberg et al., 2013). It should be noted that relationship between working memory (cognitive processing) and speech-in-noise intelligibility is less evident for younger participants than older hearing-impaired participants (Füllgrabe and Rosen, 2016). In this study, a common feature of most sentence-recognition models is that long-term language knowledge helps the participant choose the appropriate phonological and lexical candidates, thereby allowing the participant to make the correct selection step by step based on the acoustic speech characteristics of the speech signal (McClelland and Elman, 1986; Norris et al., 2016). The results of this study showed that Mandarin disyllabic recognition relies more on tone recognition, which is consistent with a bottom-up mechanism, whereas sentence recognition is inferred from context, which is consistent with the top-down mechanism of speech recognition.

We found that Region 1 + 2 has the highest score among two-frequency regions. In addition, Region 2 + 3, Region 2 + 3 + 4, and Region 1 + 2 + 4 also yielded high scores; however, Region 2 had the second-lowest relative weight. Warren et al. (1995) reported that regions centered at 370 and 6,000 Hz strongly synergize but provide little information when presented separately. Healy et al. (Healy and Warren, 2003) also found that unintelligible individual speech regions became comprehensible when combined. We assessed participant performance under all possible combinations of frequency regions and found that frequency region 2 had synergistic effects when combined with adjacent regions (i.e., frequency regions 1 or 3), which led to increased word recognition scores.

This study had some limitations. First, participants were well educated, and the influence of education level has not been evaluated in previous studies. In addition, since cognitive abilities are associated with changes in speech processing with age (even in the absence of audiometric hearing loss) (Füllgrabe et al., 2014), our findings may not be directly applicable to CI performance in different age groups. Future studies are needed to evaluate the factors of age, education level, and cognitive ability.

Overall, E cues contained in low-frequency spectral regions are more important in quiet environments for Mandarin disyllabic word recognition than for non-tonal languages such as English. These differences may be determined by the tone characteristics of Mandarin Chinese, but the influence of signal extraction parameters, test materials, and test environments cannot be excluded.

CONCLUSION

1. We found that E cues in frequency regions 1 (80–502 Hz) and 3 (1,022–1,913 Hz) significantly contributed to Mandarin disyllabic word recognition in quiet.

2. In contrast to English speech recognition, the low-frequency region contributed strongly to Mandarin Chinese disyllabic word recognition due to the tonal nature of the language.

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 author/s.

ETHICS STATEMENT

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

AUTHOR CONTRIBUTIONS

ZZ: conceptualization and writing—original draft. KL and YG: methodology and data curation. XW, LX, and CL: investigation. SH: supervision. GF: validation and project administration. YF: conceptualization, resources, writing—review and editing, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the National Natural Science Foundation of China (Grant No. 81771015), 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 all the patients and their families for supporting our work.

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

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.

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

Desroches, A. S., Newman, R. L., and Joanisse, M. F. (2009). Investigating the time course of spoken word recognition: electrophysiological evidence for the influences of phonological similarity. *J. Cogn. Neurosci.* 21, 1893–1906. doi: 10.1162/jocn.2008.21142

Dolan, R. J., Fink, G. R., Rolls, E., Booth, M., Holmes, A., Frackowiak, R. S., et al. (1997). How the brain learns to see objects and faces in an impoverished context. *Nature* 389, 596–599. doi: 10.1038/39309

Drullman, R., Festen, J. M., and Plomp, R. (1994a). Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* 95(5 Pt 1), 2670–2680. doi: 10.1121/1.409836

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

Etchepareborda, M. C. (2003). Intervention in dyslexic disorders: phonological awareness training. *Rev. Neurol.* 36, S13–S19.

Feng, Y. M., Xu, L., Zhou, N., Yang, G., and Yin, S. K. (2012). Sine-wave speech recognition in a tonal language. *J. Acoust. Soc. Am.* 131, EL133–EL138.

French, N. R., and Steinberg, J. C. (2005). Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.* 19, 90–119. doi: 10.1121/1.1916407

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

Füllgrabe, C., and Rosen, S. (2016). On the (un)importance of working memory in speech-in-noise processing for listeners with normal hearing thresholds. *Front. Psychol.* 7:1268.

Füllgrabe, C., Moore, B. C., and Stone, M. A. (2014). Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition. *Front. Aging Neurosci.* 6:347.

Füllgrabe, C., Stone, M. A., and Moore, B. C. (2009). Contribution of very low amplitude-modulation rates to intelligibility in a competing-speech task (L). *J. Acoust. Soc. Am.* 125, 1277–1280. doi: 10.1121/1.3075591

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.

Healy, E. W., and Warren, R. M. (2003). The role of contrasting temporal amplitude patterns in the perception of speech. *J. Acoust. Soc. Am.* 113, 1676–1688. doi: 10.1121/1.1553464

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.

Kim, B. J., Chang, S. A., Yang, J., Oh, S. H., and Xu, L. (2015). Relative contributions of spectral and temporal cues to Korean phoneme recognition. *PLoS One* 10:e0131807. doi: 10.1371/journal.pone.0131807

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, 2815–2824. 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.

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

McClelland, J. L., and Elman, J. L. (1986). The TRACE model of speech perception. *Cog. Psychol.* 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

McRackan, T. R., Bauschard, M., Hatch, J. L., Franko-Tobin, E., Droghini, H. R., Nguyen, S. A., et al. (2018). Meta-analysis of quality-of-life improvement after cochlear implantation and associations with speech recognition abilities. *Laryngoscope* 128, 982–990. doi: 10.1002/lary.26738

Moore, B. C. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *J. Assoc. Res. Otolaryngol.* 9, 399–406. doi: 10.1007/s10162-008-0143-x

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. *Int. J. Audiol.* 44, 379–390. doi: 10.1080/14992020500147615

Norris, D., McQueen, J. M., and Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Lang. Cogn. Neurosci.* 31, 4–18. doi: 10.1080/23273798.2015.1081703

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., et al. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Front. Syst. Neurosci.* 7:31.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 336, 367–373. doi: 10.1098/rstb.1992.0070

Rosen, S., and Hui, S. N. (2015). Sine-wave and noise-vocoded sine-wave speech in a tone language: acoustic details matter. *J. Acoust. Soc. Am.* 138, 3698–3702. doi: 10.1121/1.4937605

Shannon, R. V., Fu, Q. J., and Galvin, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Otolaryngol. Suppl.* 552, 50–54. doi: 10.1080/03655230410017562

Shannon, R. V., Galvin, J. J., 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

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

Stone, M. A., Füllgrabe, C., and Moore, B. C. (2008). Benefit of high-rate envelope cues in vocoder processing: effect of number of channels and spectral region. *J. Acoust. Soc. Am.* 124, 2272–2282. doi: 10.1121/1.2968678

Tavakoli, M., Jalilevand, N., Kamali, M., Modarresi, Y., and Zarandy, M. M. (2015). Language sampling for children with and without cochlear implant: MLU, NDW, and NTW. *Int. J. Pediatr. Otorhinolaryngol.* 79, 2191–2195. doi: 10.1016/j.ijporl.2015.10.001

Tuennerhoff, J., and Noppeney, U. (2016). When sentences live up to your expectations. *NeuroImage* 124, 641–653. doi: 10.1016/j.neuroimage.2015.09.004

Wang, S., Dong, R., Liu, D., Zhang, L., and Xu, L. (2016). The Relative contributions of temporal envelope and fine structure to Mandarin lexical tone perception in auditory neuropathy spectrum disorder. *Adv. Exp. Med. Biol.* 894, 241–248. doi: 10.1007/978-3-319-25474-6_25

Wang, S., Mannell, R., Newall, P., Zhang, H., and Han, D. (2007). Development and evaluation of Mandarin disyllabic materials for speech audiometry in China. *Int. J. Audiol.* 46, 719–731. doi: 10.1080/14992020701558511

Warren, R. M., Bashford, J. A., 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

Warren, R. M., Riener, K. R., Bashford, J. A. Jr., and Brubaker, B. S. (1995). Spectral redundancy: intelligibility of sentences heard through narrow spectral slits. *Percept. Psychophys.* 57, 175–182. doi: 10.3758/bf03206503

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

WHO (2020). *Deafness and Hearing Loss*. Available Online at: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (accessed April 1, 2021).

Wong, L. L., Ho, A. H., Chua, E. W., and Soli, S. D. (2007). Development of the Cantonese speech intelligibility index. *J. Acoust. Soc. Am.* 121, 2350–2361.

Wong, P., and Strange, W. (2017). Phonetic complexity affects children's Mandarin tone production accuracy in disyllabic words: a perceptual study. *PLoS One* 12:e0182337. doi: 10.1371/journal.pone.0182337

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., and Pfingst, B. E. (2008). Spectral and temporal cues for speech recognition: implications for auditory prostheses. *Hear Res.* 242, 132–140. doi: 10.1016/j.heares.2007.12.010

Xu, L., and Zheng, Y. (2007). Spectral and temporal cues for phoneme recognition in noise. *J. Acoust. Soc. Am.* 122:1758. doi: 10.1121/1.2767000

Xu, L., and Zhou, N. (2011). "Tonal languages and cochlear implants," in *Auditory Prostheses: New Horizons*, eds F. Zeng, A. Popper, and R. Fay (New York, NY: Springer), 341–364. doi: 10.1007/978-1-4419-9434-9_14

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

Xu, L., Tsai, Y., and Pfingst, B. E. (2002). Features of stimulation affecting tonal-speech perception: implications for cochlear prostheses. *J. Acoust. Soc. Am.* 112, 247–258. doi: 10.1121/1.1487843

Yang, J., Zhang, Y., Li, A., and Xu, L. (2017). "On the duration of Mandarin tones," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2017)*, (Stockholm).

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.