Applied Acoustics 205 (2023) 109288

Contents lists available at ScienceDirect

Applied Acoustics

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

An improved vocoder algorithm based on music harmonics and time sampling

Qiang Meng [a,1], Guoyang Liu [b,1], Lan Tian [a,*], Ming Zeng [a], Xiaoshan Lu [c], Jiameng Yan [a]

[a] School of Microelectronics, Shandong University, Jinan, China
[b] Department of Psychology, The University of Hong Kong, Hong Kong, China
[c] School of Information Science and Engineering, Shandong University, Qingdao, China

ARTICLE INFO

Article history:
Received 20 March 2022
Received in revised form 17 January 2023
Accepted 16 February 2023
Available online 20 February 2023

Keywords:
Vocoder
Musical harmonics
Pitch ranking
Melody recognition
Timbre

ABSTRACT

Music plays an essential role in the healthy life of humans. However, few sound processor algorithms effectively encode the fine structure cues of music, resulting in inferior music perception for cochlear implant (CI) users, especially in pitch and melody. In this study, an improved music vocoder algorithm is proposed based on the conjunction of harmonic and time sampling (HTS). The algorithm includes two branches: the first, the pitch (i.e., fundamental frequency, F0) and important harmonics of music signal are extracted. Second, on the existing CI channels, the music signal is split into multiple sub-bands and the relevant envelopes are respectively matched to the F0 and important lower harmonics, then modulated and aligned with appropriate intervals which are not less than the auditory nerve response absolute refractory period (ANR-ARP). The violin sounds were synthesized and experimented with the CI tone vocoder, and the HTS algorithm was compared and evaluated with the classical continuous interleaved sampling (CIS) algorithm. Twenty normal hearing (NH) subjects were recruited for audiometry experiments. The results showed that the pitch ranking scores of the HTS were obviously better than that of the CIS, and in quiet and noisy conditions the melody recognition rates of the HTS were 46.4% and 49% higher than that of the CIS, respectively. And further results showed that the HTS algorithm also increased the timbre perception of CI vocoder music. It is suggested that the HTS algorithm has the potential to enhance the music perception of CI users.

1. Introduction

Patients with severe to profound sensorineural hearing loss can partially recover their hearing with a cochlear implant (CI) [1]. In quiet situations, CI users can recognize speech sentences with an accuracy of up to 90% [2], but their music perception is not satisfactory[3,4]. For post-lingual CI users, music is usually perceived as out-of-tune, discordant, indistinct, and emotionless sounds [5]. In addition, CI users report that certain bowed instrument sounds (such as the violin and cello) are the worst for recognition [6,7].

However, music is essential for human life and the external sound processor plays an important role in the CI system, the auditory perception of CI users is mainly determined by the vocoder algorithm [8]. Music and speech signal characteristics have several differences. Traditional sound processor algorithms mainly focused on the transmission of spectral envelope features that were important for speech perception but ignored the transmission of fine structure features that were more important for music perception [9,10]. A variety of CI encoding algorithms have been studied. For example, the continuous interleaved sampling (CIS) algorithm encoded the equally spaced average amplitude of the divided band-pass filtered output signal, which can effectively convey the spectral envelope characteristics of speech [11]. The subsequently advanced combination encoder (ACE) algorithm could dynamically select stimulation electrodes and better convey the time domain changes of the filter output amplitude of each sub-band [12] by improving the stimulation rate of the channel signal. To improve pitch perception of speech and music signals, some studies proposed encoding algorithms that endeavor to transfer fine structure features, such as the harmonic single sideband encoder (HSSE) algorithm, which encoded fine structure features by frequency downshifting [13]. Another algorithm increased the transfer of fine structure features by triggering the time series at the zero crossing of the sub-bands [14]. Although these encoding algorithms improved the speech perception of CI users to a certain extent, it

* Corresponding author at: School of Microelectronics, Shandong University, 27 Shanda South Road, Jinan, Shandong 250100, China.
E-mail address: tianlansdu@foxmail.com (L. Tian).

[1] First authors contributed equally.

https://doi.org/10.1016/j.apacoust.2023.109288
0003-682X/© 2023 Elsevier Ltd. All rights reserved.

is still difficult to listen to music for most CI users, especially in pitch and timbre perception.

In fact, the basic characteristics of music signals include rhythm, melody, intensity, and timbre. In general, music signals are composed of several sounds coming from different instruments, but each instrument can be regarded as an independent sound source. To analyze the composition and characteristics of music signals, we start with a single-instrument sound signal. The sound of a single instrument is excited by a source, and the resonant cavity of the instrument remains relatively unchanged and emits the sound, so the sound forms the specific timbre. When different notes are played from the instrument, the vibration frequencies (i.e., the fundamental frequency or F0) of the excitation source are changed according to the notes. As shown in Fig. 1(a), the frequency domain transfer function of the instrument (i.e., system function) remains basically unchanged, that is, the spectrum envelope of the music signal produced by the instrument is similar, and different music notes of the instrument (i.e., music-1 and music-2 in Fig. 1(a)) is expressed by the spectrum fine structure features of the emitted sound. The actual spectrograms of notes C4 and E4 of the violin instrument are represented by music-1 and music-2 in Fig. 1 (a), respectively. The harmonics distribution (also known as the frequency domain fine structure) of different notes (or pitches) of the instrument is obviously different. Meanwhile, this difference should also correspond to the envelope variation rate of each sub-band of the CI sound processor. The fine structure is different from note to note. However, the musical note is the basic element that dominates the pitch or melody of music signals, and the total number is 120, which can be divided into 10 octaves. As shown in Table 1, each note corresponds to a fixed F0 [15]. Thus, although the existing CI products with 22 electrodes can convey the rhythm and intensity characteristics of music well, it is impossible to encode the pitch of notes completely. Even for individual instrumental sounds, this crude spectral resolution encoded with constant time sampling to convey musical note cues makes it difficult for CI users to obtain tonal or melodic perception effectively. To improve music perception of CI users, the musical pitch cues should be encoded explicitly on the existing nerve electrode pathways with an appropriate improved algorithm. Therefore, this paper introduces the optimized encoder algorithm on existing electrodes structure, which is based on harmonic time–frequency sampling (HTS) of music note.

The signal output of each sub-band of CI processor needs to be sent to each electrode of the cochlear auditory nerve pathway. Then the stimulation signal makes the neuron fire, evokes the conduction of the electrical signal of the auditory nerve, and further reconstructs the auditory perception in the cortex. However, the electrical stimulation signal must comply with nerve conduction response characteristics [16,17]. Neuro-electrophysiology has known that the auditory nerve has a phase-locking feature [18], while the harmonic cues of the music have obvious periodicity in time domains. Besides, the auditory nerve response absolute refractory period (ANR-ARP) is the minimum response duration of an auditory nerve stimulus, that is, the auditory nerve will not respond to other stimuli within an ANR-ARP [19,20]. Therefore, the electrical stimulation interval should not be less than the ANR-ARP. This time limitation must be fully considered in the proposed encoding algorithm, and the pitch feature time coding is

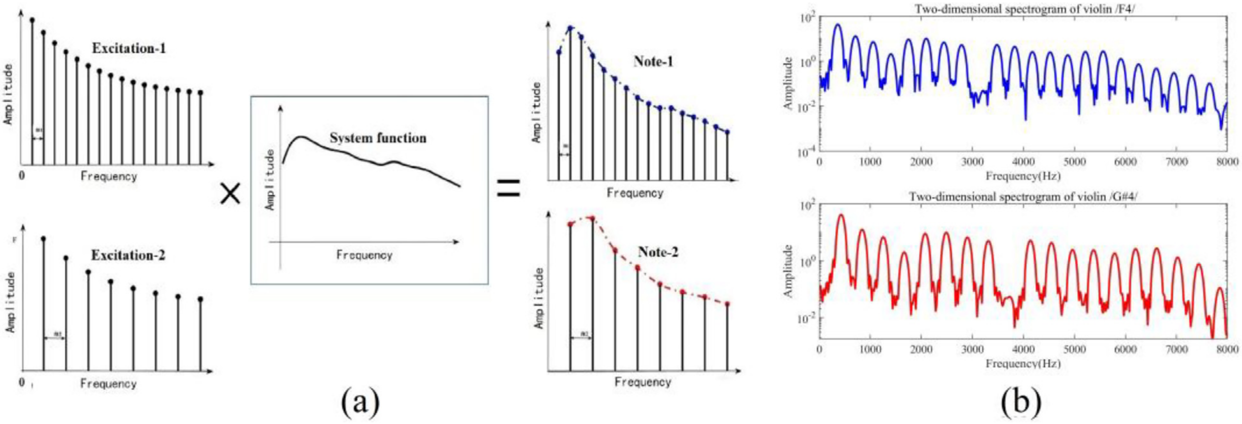


Fig. 1. (a) A schematic diagram of the spectrum structure of different music notes of the instrument. The excitation-1 corresponds to music note-1signal spectrum-1, and the excitation-2 corresponds to music note-2. The intermediate system function represents the transfer characteristic of the instrument resonant cavity. (b) The two-dimensional spectrogram of violin instrument sound /F4/ (top), /G#4/ (bottom).

Table 1
Correspondence between notes and fundamental frequencies (Hz).

Octave	C (do)	$C^{\#}/D^{b}$	D (re)	$D^{\#}/E^{b}$	E (mi)	F (fa)	$F^{\#}/G^{b}$	G (so)	$G^{\#}/A^{b}$	A (la)	$A^{\#}/B^{b}$	B (si)
0	16.3	17.3	18.4	19.4	20.6	21.8.	23.1	24.5	25.9	27.5	29.1	30.9
1	32.7	34.6	36.7	38.9	41.2	43.7	46.3	49.0	51.9	55.0	58.3	61.7
2	65.4	69.3	73.4	77.8	82.4	87.3	92.5	98.0	103.8	110.0	116.5	123.5
3	130.8	138.6	146.8	155.6	164.8	174.6	185.0	196.0	207.6	220.0	233.1	246.9
4	261.6	277.2	293.7	311.1	329.2	349.2	370.0	392.0	415.3	440.0	466.2	493.9
5	523.3	554.4	587.3	622.3	659.3	698.5	740.0	784.0	830.6	880.0	932.4	987.8
6	1046.5	1106.8	1174.7	1244.5	1318.5	1396.9	1480.0	1568.0	1661.3	1760.0	1864.7	1975.6
7	2093.1	2217.5	2349.4	2489.1	2637.1	2793.9	2960.0	3136.0	3322.5	3520.1	3729.4	3951.2
8	4186.1	4435.0	4698.8	4978.2	5274.2	5587.8	5920.1	6272.1	6645.0	7040.2	7458.8	7902.3
9	8372.2	8870.1	9397.5	9956.3	10518.3	11175.6	11840.1	12544.2	13290.1	14080.3	14917.6	15804.6

based on the rule of the nearest time alignment at the ANR-ARP interval.

Based on the above analyses, the HTS algorithm is proposed, which conveys the musical F0 and harmonic features by time–frequency conjunctive coding. Generally, the actual verification needs new designed CI product and tuning platform. However, the CI vocoder simulation was an effective method for predicting CI auditory perception [21–24]. Therefore, HTS synthetic sound can be generated by the CI vocoder. The performance of the HTS algorithm on music perception also can be evaluated by the pitch ranking and melody recognition experiments for normal hearing (NH) subjects.

2. Materials and methods

2.1. Methods

2.1.1. CIS algorithm

The CIS algorithm is a classical sound encoding algorithm used in existing CI products, which includes stages of pre-emphasis filtering, band-pass filtering, envelope detection (consists of a rectifier followed by a low-pass filter), compression, and modulation [11]. In the CIS algorithm, the amplitude envelope of each sub-band of the sound is extracted and then coded to synthesize the CI sound or to modulate the pulsed current level using a temporal constant rate. However, the pitch (i.e., F0) and the harmonic characteristics of a musical note are not focused and encoded accordingly. This might be the key reason for poor music perception of CIs.

2.1.2. HTS algorithm

Although modern CIs provide up to 22 stimulation channels, information transfer remains limited for the perception of fine spectrotemporal details [1]. Moreover, phase information also affects the performance of music perception in CIs [25]. Therefore, the HTS algorithm is based on the CIS algorithm and adds the temporal-frequency conjunctive encoding of important harmonics, including the musical F0 extraction, the nearest important harmonic mapping, the nearest pitch time alignment as well as the amplitude envelope extraction on the existing sub-bands. Fig. 2 depicts the flowchart of the HTS algorithm.

The specific steps of the HTS algorithm were as follows.

(1) The music signal was preprocessed with a window length of 20 ms, and then the per-frame signal was pre-emphasized and filtered using the first-order high-pass Butterworth filter with a cutoff frequency of 1200 Hz.
(2) An all-phase band-pass filter bank was adopted, which is an all-pass filter bank and each band-pass filter is an all-phase FIR filter with no phase distortion and steep amplitude frequency response [26,27]. The 22-channel 127-order all-pass filter bank was designed to perform band-pass filtering on the pre-emphasized signal. In this processor, the frequency band was divided according to the non-linear Mel scale, with a frequency range of 80 $\sim$ 11025 Hz.
(3) The envelope of each band signal was extracted by half-wave rectification and low-pass filtering (the fourth-order Butterworth filter with a cutoff frequency of 400 Hz).
(4) On the other hand, a 1024-point fast Fourier transform analysis was performed on each frame of the music signal after preprocessing, and the fundamental frequency of the per-frame signal was extracted by subharmonic summation [28]. Subsequently, the frequency and phase of each harmonic signal were obtained according to the multiple relationships between the fundamental frequency and the harmonics.
(5) The frequency of each band signal was selected according to the nearest harmonic mapping rule. Because it is impossible to completely transmit the harmonic characteristics of

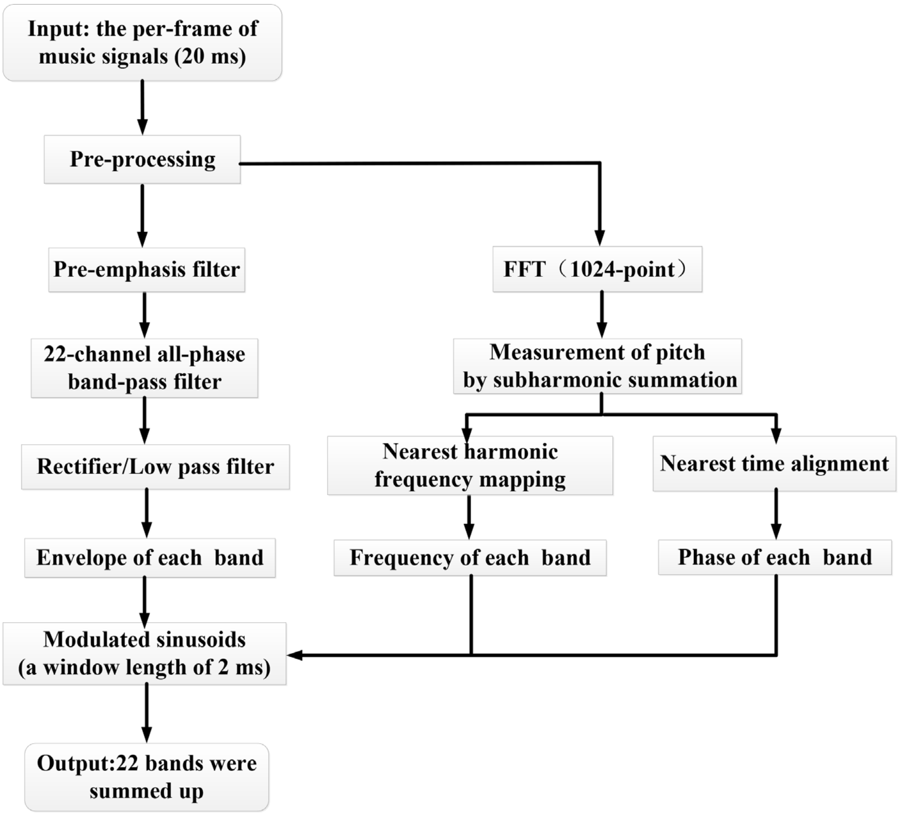

Input: the per-frame of music signals (20 ms)
Pre-processing
Pre-emphasis filter
22-channel all-phase band-pass filter
Rectifier/Low pass filter
Envelope of each band
Modulated sinusoids (a window length of 2 ms)
Output:22 bands were summed up
FFT（1024-point）
Measurement of pitch by subharmonic summation
Nearest harmonic frequency mapping
Nearest time alignment
Frequency of each band
Phase of each band

Fig. 2. Flowchart of the HTS vocoder algorithm.

music with the current 22 channels, we proposed the nearest harmonic mapping rule to transmit the important harmonics of the music signal. That is, the frequency of the lowest harmonic in each band signal was selected as the frequency of the CI vocoder's corresponding band. This means that the important harmonic frequency of each band signal was transmitted. Fig. 3(a) shows the frequency selection process of each band signal under the HTS algorithm.

(6) The phase of each band signal was selected according to the nearest time alignment rule. Considering the limitation of the ANR-ARP period, we proposed the nearest time alignment rule based on the characteristics of the auditory nerve. That is, the stimulation interval of each band signal was approximated to an integral multiple of the ANR-ARP, which is the minimum stimulation interval in CI. In the vocoder simulation, the phase of the lowest harmonic in each band signal was selected as the phase of the corresponding CI band. Therefore, the HTS algorithm can effectively encode harmonic phase features. Fig. 3(b) shows the stimulation interval selection process by the HTS algorithm.

(7) Finally, the envelope of each band signal was used to modulate a sinusoid carrier at the respective band frequency and phase. The modulated sinusoid signals were then combined across all bands to generate the HTS synthesized sound; the synthetic window length of the sound was 2 ms.

2.2. Experiments and materials

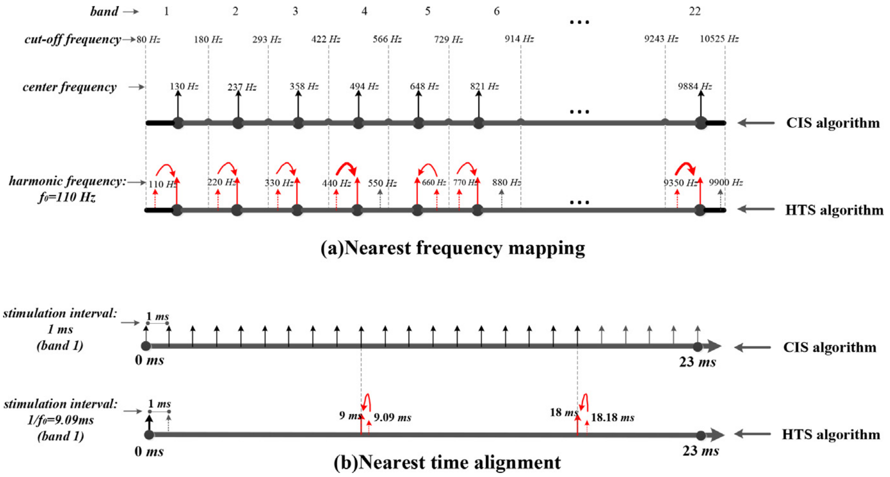
To investigate the performance of the HTS algorithm in musical pitch perception, we conducted pitch ranking and melody recognition experiments. The vocoder simulation platform was MATLAB 2020a. All experiment stimuli sounds come from violin, because the violin is a typical stringed instrument with abundant harmonics and a complex vibration system [31], and its sound is the worst for CI users, then the violin-based experiments have practical value. The violin sound was created by the Composer Master software [32], and the sampling frequency was 22.05 kHz.

In studies of CI speech coding algorithms, two vocoder types were often used: noise vocoder and tone vocoder [21,22,33]. Because the musical note signal is all voiced sound and mainly composed of the F0 and harmonic components, the tone vocoder is more suitable in CI music simulation [34]. Moreover, the pulse-spreading harmonic complex vocoder [24] is also a new type of tone vocoder. In this paper, the CI-synthesized music also used a tone vocoder. Based on existing electrode bands (i.e., 22 channels) in the HTS tone vocoder, the effect of the current spread of future CI products should be similar to that of the existing CI product. The HTS-synthesized sound was obtained according to the algorithm flow in Fig. 2. While the CIS synthesized sound was obtained only by the envelope modulation of each sinusoidal carrier (i.e., tone vocoder), in which the frequency of the sinusoidal signal was the center frequency of each band [33,34], most of them do not match and align with the F0 or harmonics of the musical note.

Twenty NH subjects (twelve males and eight females, aged $20 \sim 30$, mean age = 24) were recruited to participate in this study. All subjects had no history of auditory nerve or hearing injury. Before the experiment, the potential candidates were informed of the experiment content and precautions, and they agreed to participate in the experiments. All testing was performed in an audio laboratory with a comfortable environment and without noise. The stimuli were played at a comfortable listening level via headphones (Bose QC25), and there was no feedback in the experiment.

The HTS algorithm was compared to the other two algorithms. The first algorithm was the traditional CIS algorithm (default low-pass filter (LPF) cutoff frequency of 400 Hz). The second algorithm is called the frequency mapping (FM) algorithm, which only maps important harmonics to related sub-bands without the time alignment. The FM algorithm is a middle method designed to compare. The FM-synthesized sound was obtained by the envelope modulation of the sinusoidal carrier at an important harmonic frequency of each sub-band.

In the pitch ranking experiment, subjects were asked to determine whether the final note was higher or lower in pitch than the first two and to select their response from the closed set of two options (higher and lower) on the screen in front of them [35]. Each stimulus consisted of three notes presented sequentially. Every note had a duration of 500 ms, and the time interval between notes was 240 ms. The first two notes had the same pitch, while the third note had a different pitch. A total of 450 synthesized sounds (three types of algorithms × three types of base frequencies × 5 types of semitone spacing × 10 trials) were evaluated. The base frequencies were E3 (164 Hz), C4 (262 Hz) and E5



(a)Nearest frequency mapping

(b)Nearest time alignment

Fig. 3. (a) Schematic diagram of nearest frequency mapping. If there were two or more harmonic frequencies in a certain frequency band, we discarded the higher harmonics and only kept the lowest harmonic frequencies [29,30]. If there was only one harmonic frequency in a certain frequency band, the harmonic frequency was directly mapped to the center frequency of the corresponding frequency band. When there was no harmonic signal in the band, the HTS algorithm did not perform signal processing. (b) Schematic diagram of nearest time alignment. It was assumed that when the F0 of the music signal was 110 Hz and the ANR-ARP was 1 ms, the harmonic period of the first frequency band signal was 9.09 ms, therefore the minimum stimulus interval in the frequency band was set to 9 ms. In the vocoder simulation, the phase of the frequency band signal was the phase of the 110 Hz harmonic.

(663 Hz). Eighteen notes were selected, they are 164, 174, 185, 196, 220, 246, 262, 277, 293, 311, 349, 392, 663, 698, 740, 784, 880 and 987 Hz. The stimuli were presented in pairs of pitches ranging in interval sizes of 1, 2, 3, 5, or 7 semitones, and 10 trials were presented for each interval size at each base frequency. In the experiment, the subjects were first required to obtain a score of 80% correct or higher for the pitch ranking of the original sound. Then, the synthesized sounds were divided into nine groups (three methods × three base frequencies). In the test, each synthesized sound of the group was randomly played in order to eliminate the effect of sound order on the experimental results.

In the melody recognition experiment, the subjects were asked to identify and select the melody names from the closed set of five melody names on the screen in front of them [36]. All melodies were created in middle C (the F0 range from 261 to 523 Hz). To eliminate the interference of rhythm information on melody recognition, each melody comprised 12 equal-duration notes, and the duration of each melody was about 5 s. To evaluate the robustness of the melody perception of the proposed method, we further carried out the experiments under noise conditions. A total of 30 synthesized sounds (five melodies × three encoding algorithms × two conditions (in quiet and signal–noise ratios SNRs = 0 dB)) were evaluated. Among them, the stimuli materials included five familiar music melodies, namely Happy Birthday to You, Jingle Bells, Little Tigers, Welcome to Beijing, and Jasmine. The noise type was Gaussian white noise. Before the experiment, subjects were familiarized with five original melody segments by repeat playing. Then, the experiments were divided into six groups, randomly playing the synthesized sounds of three methods (CIS, FM, HTS) in quiet or noisy (SNR = 0 dB) conditions, respectively.

3. Result

3.1. Pitch ranking

Fig. 4 shows the results for pitch ranking from vocoder synthesized sounds with the three algorithms, three base frequencies, and five semitone spacing, respectively. In general, subjects' percent correct in pitch ranking gradually increases as the semitone spacing increases. Under the three base frequency conditions, subjects' average percent correct in pitch ranking of the HTS synthesized sounds is higher than that of the other two synthesized sounds. The average percent correct of FM algorithm is 10.3%, 3.9%, and 7.3% higher than the traditional CIS algorithm at the baseline of 164 Hz, 262 Hz and 663 Hz, respectively. The results of the analysis of variance (ANOVA) suggest that the HTS algorithm significantly improves the performance of pitch ranking at the base frequencies of 663 Hz ($F_{2,12} = 4.79, p < 0.05$). However, its improvement at the base frequency of 164 Hz ($F_{2,12} = 3.17, p = 0.0783$) and 262 Hz ($F_{2,12} = 1.93, p = 0.1877$) is not significant.

3.2. Melody recognition

Fig. 5 shows the results for the melody recognition of synthesized sounds by three vocoder algorithms under two conditions (in quiet and in noisy with SNR = 0 dB). It can be seen that the average accuracy of melody recognition of the HTS is obviously higher than that of the two others. The accuracy of FM algorithm is only 6.7% and 8.3% higher than the traditional CIS algorithm in quiet and noisy, respectively. ANOVA suggests that the HTS algorithm

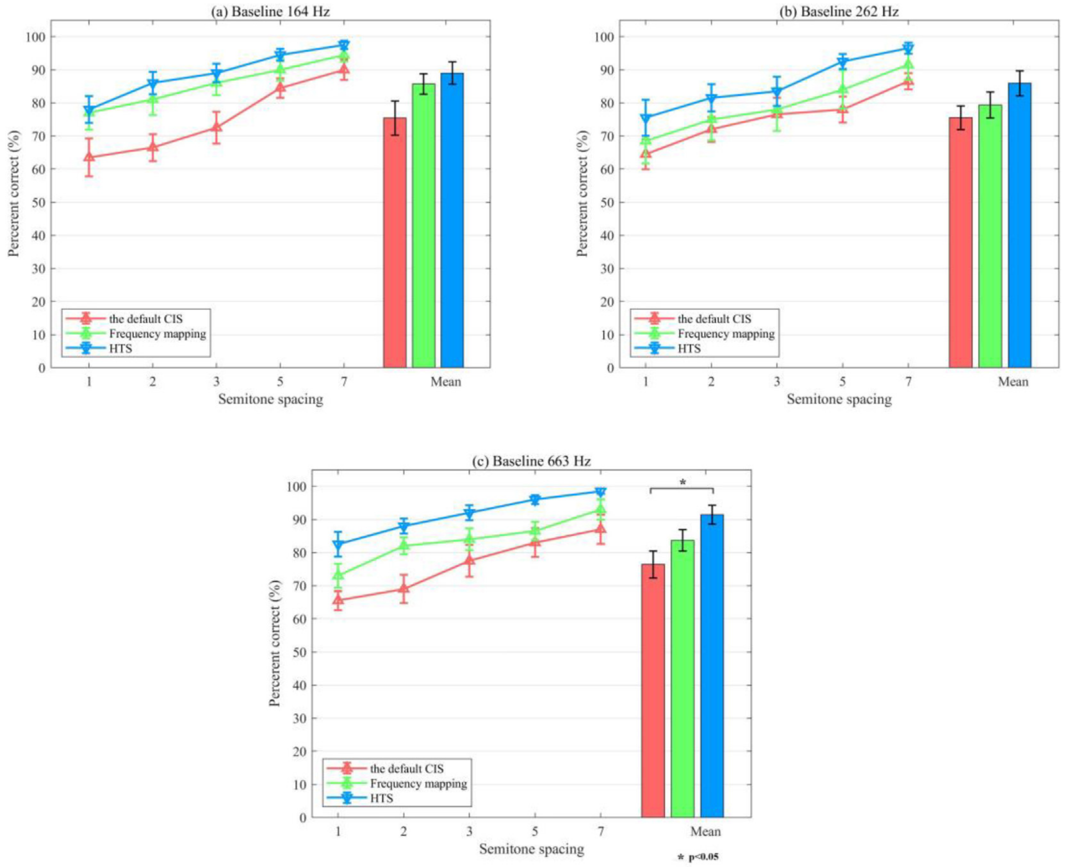
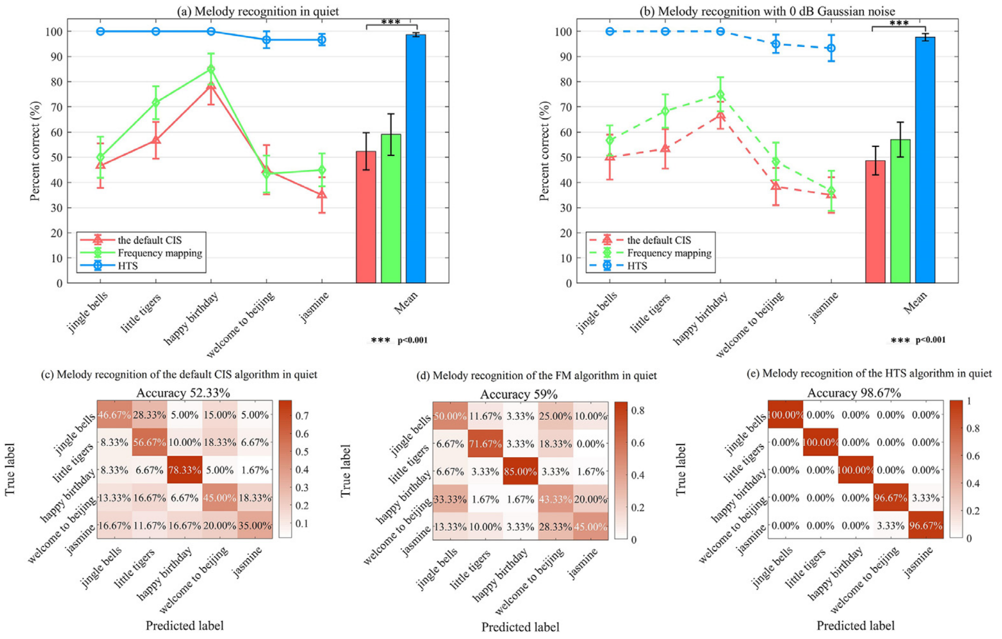


Fig. 4. Pitch ranking results under five semitone spacing conditions and three algorithms. (a) baseline 164 Hz, (b) baseline 262 Hz, (c) baseline 663 Hz. Black bars represent the standard error of the mean. The asterisk represents the significance of the analysis of variance.



(c) Melody recognition of the default CIS algorithm in quiet

Accuracy 52.33%

	jingle bells	little tigers	happy birthday	welcome to beijing	jasmine
jingle bells	46.67%	28.33%	5.00%	15.00%	5.00%
little tigers	8.33%	56.67%	10.00%	18.33%	6.67%
happy birthday	8.33%	6.67%	78.33%	5.00%	1.67%
welcome to beijing	13.33%	16.67%	6.67%	45.00%	18.33%
jasmine	16.67%	11.67%	16.67%	20.00%	35.00%

(d) Melody recognition of the FM algorithm in quiet

Accuracy 59%

	jingle bells	little tigers	happy birthday	welcome to beijing	jasmine
jingle bells	50.00%	11.67%	3.33%	25.00%	10.00%
little tigers	6.67%	71.67%	3.33%	18.33%	0.00%
happy birthday	6.67%	3.33%	85.00%	3.33%	1.67%
welcome to beijing	33.33%	1.67%	1.67%	43.33%	20.00%
jasmine	13.33%	10.00%	3.33%	28.33%	45.00%

(e) Melody recognition of the HTS algorithm in quiet

Accuracy 98.67%

	jingle bells	little tigers	happy birthday	welcome to beijing	jasmine
jingle bells	100.00%	0.00%	0.00%	0.00%	0.00%
little tigers	0.00%	100.00%	0.00%	0.00%	0.00%
happy birthday	0.00%	0.00%	100.00%	0.00%	0.00%
welcome to beijing	0.00%	0.00%	0.00%	96.67%	3.33%
jasmine	0.00%	0.00%	0.00%	3.33%	96.67%

Fig. 5. Melody recognition results. (a) the bar chart of three algorithms in quiet, (b) the bar chart of three algorithms in noisy. Black bars represent the standard error of the mean. The asterisk represents the significance of the ANOVA. And the corresponding confusion matrix of (c) the default CIS, (d) the FM, and (e) the HTS in quiet condition, respectively.

significantly improves the melody recognition performance both in quiet ($F_{2,12} = 15.35, p < 0.001$) and noise ($F_{2,12} = 25.39, p < 0.001$) conditions. The accuracy of each melody under the HTS algorithm all exceeds 90%, but the accuracies of melody under the CIS and FM algorithms are lower and unstable. Subjects are more easily able to recognize melodies when listening to the HTS synthesized music than listening to the CIS synthesized music. From Fig. 5, the results of the CIS seem quite robust to the noise, but the accuracy of melody recognition is much lower than that of the HTS. Furthermore, the melody recognition accuracy for CIS algorithm and HTS algorithm decreases by 3.67% and 1% under noise compared to the quiet condition, respectively. It turns out that the CIS algorithm has worse performance on music coding, even in quiet conditions, CIS synthesized music sounds as poor as in noisy conditions. Meanwhile, Fig. 5(c), (d) and (e) show the confusion matrix of the default CIS, FM, and HTS algorithms under quiet conditions, respectively. It can be found that the confusion matrix results match the bar chart results.

It is worth mentioning that the performance of melody recognition of the HTS is both significantly higher in quiet and noisy (SNR = 0 dB) conditions and is also stable, this indicates that the HTS algorithm has an excellent encoding effect for music features and has a strong noise robustness.

4. Discussion

4.1. Pitch

The music signal is characterized by the regular F0 distribution and harmonic interrelation. In terms of pitch recognition, the HTS algorithm performs better than two other algorithms (see Fig. 4). The main reason is that the HTS vocoder encoded the features of the F0 and some important harmonics of the note onto the corresponding sub-bands and phase timing. However, in the CIS vocoder, the central frequency of each band is fixed and mostly is not matched to the important harmonic of the note and there is also no time encoding of the music note on each band. Thus, the CIS vocoder cannot accurately transmit the music pitch and harmonic features. Moreover, we compare the original signal with the reconstructed signals using the HTS algorithm and the CIS algorithm, such as the second sub-band (168–266 Hz). As shown in Fig. 6 (a) and (b), the HTS algorithm retains more time–frequency fine cues of the music signal and the HTS vocoder reconstructed signal better retains the phase information of the original signal. In contrast, the CIS reconstructed signal is significantly inconsistent with the original signal. In addition, the overall differences between different algorithms can also be seen from the spectrogram. As shown in Fig. 7, the harmonic structure features of the HTS synthesized sound is more obvious than that of CIS synthesized sound. Moreover, the spectrogram of the HTS simulated sound is closer to the spectrogram of the original sound. Consequently, it is easy to identify the pitch ranking of the HTS-synthesized sound but not the CIS-synthesized sound.

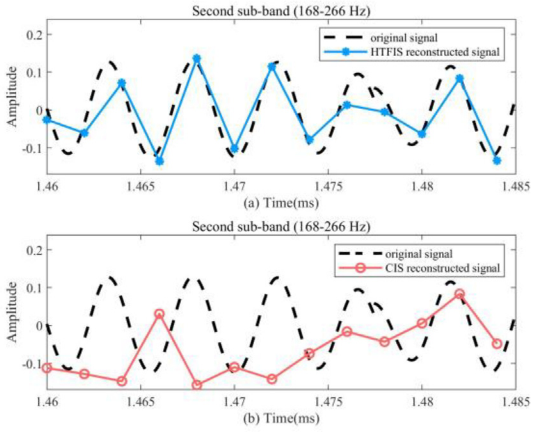


Fig. 6. Waveform of the second band of the violin note at 220 Hz. Comparison of the original signal with the reconstructed signal under the HTS algorithm (top) and the CIS algorithm (bottom).

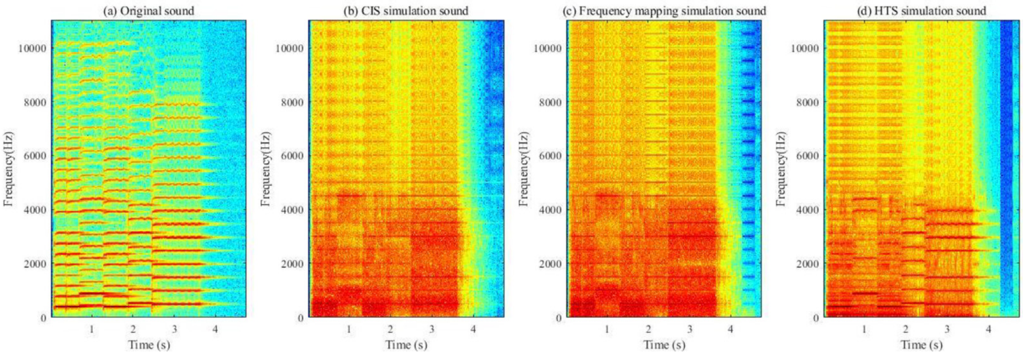


Fig. 7. The spectrograms of the original and synthesized sounds for the music Happy Birthday to You by different algorithms.

Meanwhile, the HTS algorithm is different from the F0mod strategy [37]. For the F0mod algorithm, although the input sound signal is presented to two parallel blocks: the first block is the filter banks which split the signal into 22 channels and extract the relatively slow varying envelopes of each channel. The second block estimates the F0 of the input signal, and the 22 sub-band envelopes are then modulated by only F0 sinusoidal signal [37]. In contrast, the HTS algorithm not only extracts F0 feature and the envelope of the sub-bands, but also modulates the envelope of important harmonics sub-bands of music signal by the combination coding of harmonic frequency mapping and time alignment (i.e., the phase information of the harmonic frequency). Therefore, the HTS algorithm may convey more musical features than F0mod algorithm.

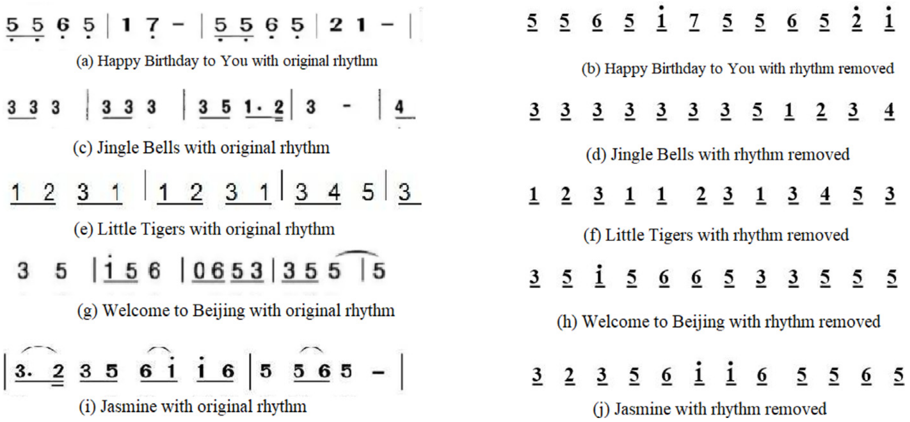
A melody is composed of multiple notes and its note changes are regular. Consequently, melody recognition should be easier than pitch ranking. From Fig. 4 and Fig. 5, it can be found that the melody recognition improvement by the HTS algorithm is greater than pitch ranking. Even under noisy condition (see Fig. 5), the results of melody recognition by the HTS algorithm are also good. This further indicates that the HTS algorithms can robustly encode pitch features for practical music segments.

Additionally, the performance of the melody Happy Birthday to You is significantly higher than the other four melodies. There may be two possible reasons for the results. First, the selected 12 notes with the same length (i.e., removed the original rhythm) (refer: Fig. 8 (a)/(b)) of the melody Happy Birthday to You are four complete melody fragments, while the 12 notes of the same duration of the other four melodies are not complete melody fragments in the end (refer: Fig. 8, (c)/(d), (e)/(f), (g)/(h), (i)/(j)). On the other hand, it may be due to the fact that the subjects are more familiar with the melody Happy Birthday to You than the other four melodies.

4.2. Timbre

Based on the existing 22 electrode channels, the HTS vocoder algorithm mainly aimed at pitch features encoding of CI synthesis music. Meanwhile, the improvement might also affect the musical quality (i.e., timbre) of CI vocoder. However, timbre is complex and abstract and has been considered one of the most difficult acoustic features to comprehend. Thus, its evaluation should be a multidimensional and complex assessment problem [38]. It is known that the perceptual evaluation of speech quality (PESQ) is a usual objective index mainly used for speech, whose score ranges from −0.5 to 4.5 [39], while the perception model-based quality (PEMO-Q) score is a more general objective index of audio quality (including speech and music), whose score ranges from −4 (i.e., very annoying impairment) to 0 (i.e., imperceptible impairment) [40]. Therefore, we use the two indexes to evaluate the timbre perception of the synthesized 18 single notes of three vocoders. In the timbre evaluation, the materials are the same as the sounds of pitch ranking test. When the testing music segments are all the same in intensity, length (i.e., no rhythm) and pitch, the scores of PESQ and PEMO-Q are calculated, respectively. Fig. 9 shows the average PESQ and PEMO-Q scores.

It can be found that the average PESQ and PEMO-Q scores of HTS algorithm are both higher than that of the other two algorithms, which indicates that the waveform distortion of the HTS


(a) Happy Birthday to You with original rhythm

(b) Happy Birthday to You with rhythm removed

(c) Jingle Bells with original rhythm

(d) Jingle Bells with rhythm removed

(e) Little Tigers with original rhythm

(f) Little Tigers with rhythm removed

(g) Welcome to Beijing with original rhythm

(h) Welcome to Beijing with rhythm removed

(i) Jasmine with original rhythm

(j) Jasmine with rhythm removed

Fig. 8. Comparison of 12 notes of the 5 melody segments.

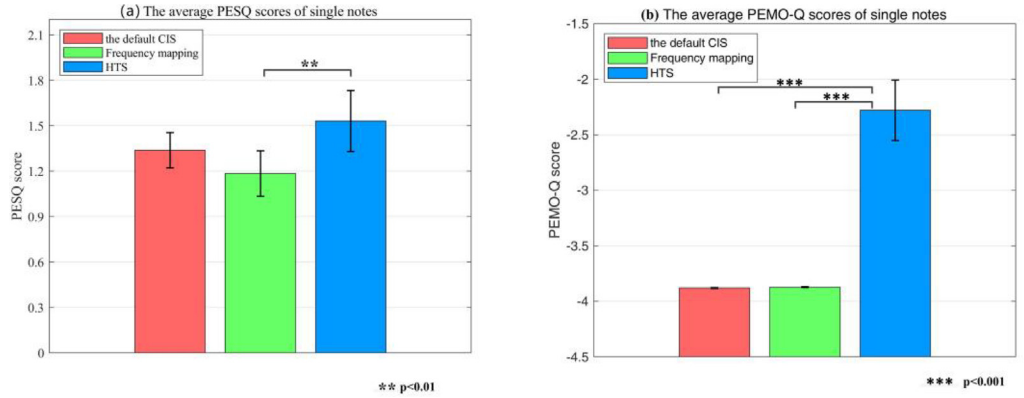


Fig. 9. (a) PESQ and (b) PEMO-Q scores of single notes of the three algorithms. Black bars represent the standard error of the mean.

synthesized music is the smallest. It is also found that the average PEMO-Q scores of HTS algorithm are much higher than that of the other two algorithms. This further indicates that the timbre degradation of the HTS-synthesized music is much smaller. Hence, the results show that the HTS vocoder algorithm not only improves CI music pitch perception but also increases the CI timbre effect. The *t*-test results show marginally significant differences for PESQ evaluation ($p = 0.074$) and significant differences for PEMO-Q evaluation ($p < 0.001$) between the default CIS and HTS.

4.3. Electrodogram

As a visual comparison, the electrodograms obtained by CIS and HTS for a violin melody of Happy Birthday to You are provided in Fig. 10(a) and (b), respectively, along with the melody's fundamental frequency in Fig. 10(c). Electrodograms are generated based on the CCi-Mobile platform [41]. The electrodogram is defined as electric pulse stimulation patterns delivered by the sound coding strategy, i.e., the current levels delivered by each electrode over time [1]. Electrodograms are an effective way to compare how acoustic features are transmitted by different sound coding strategies [13]. There are 22 electrodes in the tone vocoder simulation, while not all electrodes are active in the HTS algorithm because some lower sub-bands might not include any F0 and harmonics of the note segment. At the same time, the selected lowest electrodes (i.e. sub-bands, in Fig. 10(b)) by the HTS are consistent with the distribution of fundamental frequency in Fig. 10(c). The obvious difference between the two electrodograms also can prove the improved effect for CI synthesized music by HTS algorithm.

In addition, there are some limitations to this study. Although the performance of HTS algorithm for NH subjects has significantly improved, this is only the first step of verification work, and it may be seen as the ideal level for future CI users. Further verification of the actual product will depend on new designs on CI product software [42], decoding chips and wireless transmission protocols. Although the work will be time-consuming, it will be a practical benefit to a larger number of CI users. In this paper, this algorithm has experimented with only single instrument sound. However, for the complex sounds of multi-instruments, it could be adaptable to apply by adding the related preprocesses, such as sound source separation [43] and extraction of the main melody pitch [44]. Since the HTS algorithm only encoded and transmitted the cues of F0 and lower important harmonics of the appropriate music note, it resulted in that the higher the note, the fewer harmonics be matched and encoded on the fewer sub-bands (channels). When the pitch of a note is so high that its 1/F0 duration is less than the ANR-ARP, the HTS algorithm will become similar to the CIS or FM algorithm. Because in this case, even for the F0 component of the note, there are no bands (i.e., electrodes) to realize the enhanced time encoding, the whole harmonics of the note will be mapped to the spectral envelopes of the classical CIS algorithm. Therefore, the HTS algorithm is a partial music enhancement algorithm based on the existing CI electrode layout.

In the future CI product, the HTS algorithm may encode and convey more pitch and harmonic features of music notes. By switching speech/music mode, the HTS algorithm can extend the CI product's function based on existing electrodes to improve CI user's music perception.

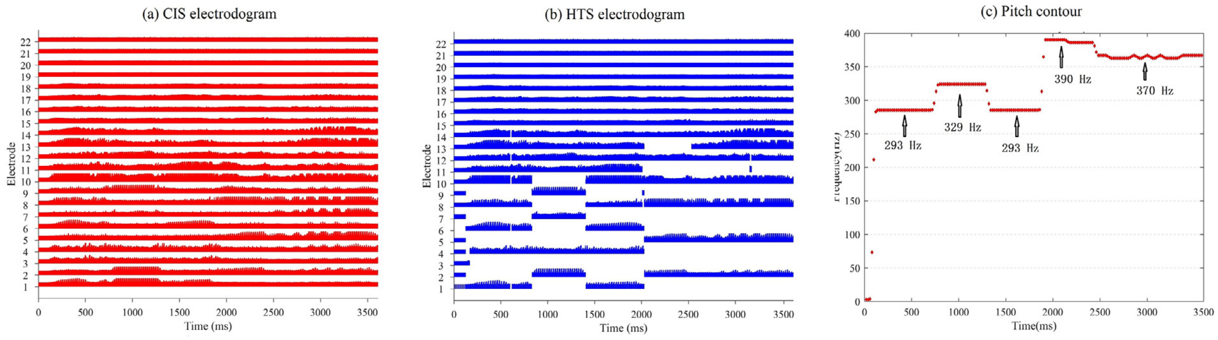


Fig. 10. The CI encoding electrodograms and pitch contour of Happy Birthday to You.

5. Conclusion

This study proposed the HTS algorithm that conjunctly encoded the frequency and temporal features of important music harmonics based on the existing CI electrode layout. We conducted pitch ranking and melody recognition experiments based on CI vocoder and compared them with CIS algorithm to evaluate the HTS effects. The experimental results showed that HTS algorithm could effectively encode and convey more cues of music features and make the pitch ranking and melody recognition significantly improved. Meanwhile, it also increased the timbre perception. Overall, the HTS algorithm has the potential to improve the music perception of CI users.

Financial disclosures/conflicts of interest

The authors have declared that no conflicts of interest, financial, or otherwise.

CRediT authorship contribution statement

Qiang Meng: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Guoyang Liu:** Data curation, Writing – review & editing. **Lan Tian:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Ming Zeng:** Investigation, Software. **Xiaoshan Lu:** Software, Validation, Data curation, Writing – review & editing. **Jiameng Yan:** Data curation, Visualization.

Data availability

Data will be made available on request.

Declaration of Competing Interest

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

Acknowledgments

This work was supported by the Natural Science Foundation of Shandong Province [grant numbers ZR2021ZD40, ZR2021MF065]; the Research Project for Graduate Education and Teaching Reform, Shandong University, China [grant number XYJG2020108].

References

[1] Nogueira W, Nagathil A, Martin R. Making music more accessible for cochlear implant listeners: Recent developments. IEEE Signal Process Mag 2019;36:115–27. https://doi.org/10.1109/MSP.2018.2874059.

[2] Wilson BS, Dorman MF. Cochlear implants: A remarkable past and a brilliant future. Hear Res 2008;242:3–21. https://doi.org/10.1016/j.heares.2008.06.005.

[3] Kalathottukaren RT, Purdy SC, Ballard E. Prosody perception and musical pitch discrimination in adults using cochlear implants. Int J Audiol 2015;54:444–52. https://doi.org/10.3109/14992027.2014.997314.

[4] Wouters J, McDermott HJ, Francart T. Sound Coding in Cochlear Implants: From electric pulses to hearing. IEEE Signal Process Mag 2015;32:67–80. https://doi.org/10.1109/MSP.2014.2371671.

[5] Jiam Nicole T, Caldwell Meredith T, Limb Charles J. What does music sound like for a cochlear implant user? Otol Neurotol 2017;38:E240–7. https://doi.org/10.1097/MAO.0000000000001448.

[6] Prevoteau C, Chen SY, Lalwani AK. Music enjoyment with cochlear implantation. Auris Nasus Larynx 2018;45:895–902. https://doi.org/10.1016/j.anl.2017.11.008.

[7] Zhang F, Benson C, Cahn SJ. Cortical encoding of timbre changes in cochlear implant users. J Am Acad Audiol 2020;24:46–58. https://doi.org/10.3766/jaaa.24.1.6.

[8] Seldran F, Gallego S, Thai-Van H, Berger-Vachon C. Influence of coding strategies in electric-acoustic hearing: A simulation dedicated to EAS cochlear implant, in the presence of noise. Appl Acoust 2014;76:300–9. https://doi.org/10.1016/j.apacoust.2013.08.003.

[9] Fischer T, Schmid C, Kompis M, Mantokoudis G, Caversaccio M, Wimmer W. Effects of temporal fine structure preservation on spatial hearing in bilateral cochlear implant users. J Acoust Soc Am 2021;150:673–86. https://doi.org/10.1121/10.0005732.

[10] Moshgelani F, Parsa V, Allan C, Veeranna SA, Allen P. Perceptual and objective assessment of envelope enhancement for children with auditory processing disorder. IEEE Trans Neural Syst Rehabil Eng 2020;28:143–51. https://doi.org/10.1109/TNSRE.2019.2957230.

[11] Wilson BS, Finley CC, Lawson DT, Wolford RD, Eddington DK, Rabinowitz WM. Better speech recognition with cochlear implants. Nature 1991;352(6332):236–8.

[12] Patrick JF, Busby PA, Gibson PJ. The development of the nucleus freedom™ cochlear implant system. Trends Amplif 2006;10:175–200. https://doi.org/10.1177/1084713806296386.

[13] Li X, Nie K, Imennov NS, Rubinstein JT, Atlas LE. Improved perception of music with a harmonic based algorithm for cochlear implants. IEEE Trans Neural Syst Rehabil Eng 2013;21:684–94. https://doi.org/10.1109/TNSRE.2013.2257853.

[14] Riss D, Hamzavi JS, Blineder M, Honeder C, Ehrenreich I, Kaider A, Baumgartner WD, Gstoettner W, Arnoldner C. FS4, FS4-p, and FSP: A 4-month crossover study of 3 fine structure sound-coding strategies. Ear Hear 2014;35:E272–81. https://doi.org/10.1097/AUD.0000000000000063.

[15] Berezovsky J. The structure of musical harmony as an ordered phase of sound: A statistical mechanics approach to music theory. Sci Adv 2019;5:1–8. https://doi.org/10.1126/sciadv.aav8490.

[16] Hughes ML, Laurello SA. Effect of stimulus level on the temporal response properties of the auditory nerve in cochlear implants. Hear Res 2017;351:116–29. https://doi.org/10.1016/j.heares.2017.06.004.

[17] Limb CJ, Roy AT. Technological, biological, and acoustical constraints to music perception in cochlear implant users. Hear Res 2014;308:13–26. https://doi.org/10.1016/j.heares.2013.04.009.

[18] Taberner AM, Charles Liberman M. Response properties of single auditory nerve fibers in the mouse. J Neurophysiol 2005;93:557–69. https://doi.org/10.1152/jn.00574.2004.

[19] Goldwyn JH, Rubinstein JT, Shea-Brown E. A point process framework for modeling electrical stimulation of the auditory nerve. J Neurophysiol 2012;108:1430–52. https://doi.org/10.1152/jn.00095.2012.

[20] Miller CA, Abbas PJ, Robinson BK. Response properties of the refractory auditory nerve fiber. J Assoc Res Otolaryngol 2001;2:216–32. https://doi.org/10.1007/s101620010083.

[21] Chen Fei, Lau AdaHY. Effect of vocoder type to Mandarin speech recognition in cochlear implant simulation. In: 9th International Symposium on Chinese Spoken Language Processing (ISCSLP). p. 551–4. https://doi.org/10.1109/ISCSLP.2014.6936705.

[22] Mehta AH, Oxenham AJ. Vocoder simulations explain complex pitch perception limitations experienced by cochlear implant users. J Assoc Res Otolaryngol 2017;18:789–802. https://doi.org/10.1007/s10162-017-0632-x.

[23] Friesen LM, Shannon RV, Baskent D, Wang X. Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. J Acoust Soc Am 2001;110:1150–63. https://doi.org/10.1121/1.1381538.

[24] Karoui C, James C, Barone P, Bakhos D, Marx M, Macherey O. Searching for the sound of a cochlear implant: Evaluation of different vocoder parameters by cochlear implant users with Single-Sided deafness. Trends Hear 2019;23:1–15. https://doi.org/10.1177/2331216519866029.

[25] Sit J-J, Simonson AM, Oxenham AJ, Faltys MA, Sarpeshkar R. A Low-Power asynchronous interleaved sampling algorithm for cochlear implants that encodes envelope and phase information. IEEE Trans Biomed Eng 2007;54:138–49. https://doi.org/10.1109/TBME.2006.883819.

[26] Huang Xiangdong, Wang Zhaohua. FIR filter design based on all-phase amplitude-frequency characteristic compensation. J Circuits Syst 2008:1–5.

[27] Meng Qiang, Tian Lan, Xu Dongping, Lu Xiaoshan. Research on parameter optimization of all-phase band-pass filter for auditory reconstruction. J Fudan Univ (Nat Sci) 2020;59:551–7.

[28] Hermes DJ. Measurement of pitch by subharmonic summation. J Acoust Soc Am 1988;83:257–64. https://doi.org/10.1121/1.396427.

[29] Nemer JS, Kohlberg GD, Mancuso DM, Griffin BM, Certo MV, Chen SY, et al. Reduction of the harmonic series influences musical enjoyment with cochlear implants. Otol Neurotol 2017;38:31–7. https://doi.org/10.1097/MAO.0000000000001250.

[30] Ma Fajiang, Bai Geng Tian. Study on the Relationship between Melody Perception and Music Harmonics. AER-Adv Eng Res 2017;86:83–5.

[31] Silvestri P, Ravina E. On the vibro-acoustic characterization of two similar violas da gamba. Appl Acoust 2021;177:. https://doi.org/10.1016/j.apacoust.2021.107963107963.

[32] Composer Master Music Dreamer Software manual, Fengya Software Company; 2019.

[33] Lorenzi C, Gilbert G, Carn H, Garnier S, Moore BCJ. Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. Proc Natl Acad Sci USA 2006;103(49):18866–9.

[34] Crew Joseph D, Galvin John J. Channel interaction limits melodic pitch perception in simulated cochlear implants. J Acoust Soc Am 2012;132: EL429–35. https://doi.org/10.1121/1.4758770.

[35] Gfeller K, Turner C, Oleson J, Zhang X, Gantz B, Froman R, et al. Accuracy of cochlear implant recipients on pitch perception, melody recognition, and speech reception in noise. Ear Hear 2007;28:412–23. https://doi.org/10.1097/AUD.0b013e3180479318.

[36] Parkinson Aaron J, Rubinstein Jay T, Drennan Ward R, Dodson Christa, Nie Kaibao. Hybrid music perception outcomes: Implications for melody and timbre recognition in cochlear implant recipients. Otol Neurotol 2019;40: E283–9. https://doi.org/10.1097/MAO.0000000000002126.

[37] Laneau J, Wouters J, Moonen M. Improved music perception with explicit pitch coding in cochlear implants. Audiol Neurotol 2006;11:38–52. https://doi.org/10.1159/000088853.

[38] Wei Y, Gan L, Huang X. A Review of Research on the Neurocognition for Timbre Perception. Front Psychol 2022;13:1–9. https://doi.org/10.3389/fpsyg.2022.869475.

[39] Ma J, Hu Y, Loizou PC. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J Acoust Soc Am 2009;125:3387–405. https://doi.org/10.1121/1.3097493.

[40] Huber R, Kollmeier B. PEMO-Q—a new method for objective audio quality assessment using a model of auditory perception. IEEE Trans Audio Speech Lang Process 2006;14:1902–11. https://doi.org/10.1109/TASL.2006.883259.

[41] Ghosh R, Ali H, Hansen J. CCi-MOBILE: A portable real time speech processing platform for cochlear implant and hearing research. IEEE Trans Biomed Eng 2022;69:1251–63. https://doi.org/10.1109/TBME.2021.3123241.

[42] Tian Lan, Meng Qiang, Li Meng, Li Weiqi, and Han Xiao, An optimized coding method and system for enhancing pitch perception in cochlear implants; 2018.

[43] Bi Chuan-Xing, Chen Xin-Zhao, Chen Jian. Sound field separation technique based on equivalent source method and its application in nearfield acoustic holography. J Acoust Soc Am 2008;123:1472–8. https://doi.org/10.1121/1.2837489.

[44] Liang S, Shu R, Kaifa Z. Extraction of music main melody and Multi-Pitch estimation method based on support vector machine in big data environment. J Environ Public Health 2022;2022:1–11.