RESEARCH ARTICLE

Is Time Theory Necessary to Answer Resolved and Unresolved Harmonics Problems in Pitch Perception?

Jun-ichi Takahashi

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

The problem of pitch perception has been the subject of a long debate between place theory and time theory. Here, we propose a Power Series Template (PoST) model to answer the questions of how and why pitch perception comes about. The sensitive measurement of acoustic signals requires efficient sound amplification, which inevitably accompanies mode coupling due to the perturbative non-linearities. Under reinforcement learning, with the second- and third-order nonlinearities as default teacher signals for sound localization in auditory scene analysis, a chain of coincidence is generated. After learning is completed, two power series templates of 2^n and 3^m are generated, and the $2f_1 - f_2$ coupling of the elements contained therein fills the blanks of matchable harmonics up to N=10 in the template, providing the well-known harmonic template. When complex tones containing multiple harmonics are input into the trained network, two power series are evoked from each harmonic, and from their intersection, the brain acquires the fundamental of the complex tone as the pitch. On the other hand, this harmonic template has deficiencies for N>10. These deficiencies result in the deterioration of the fundamental discrimination threshold (F0DT) that appears in the pitch perception of harmonic complex tones above the 10th order. Based on this template model, consistent explanations are given for the problems of the missing fundamental, octave equivalence, pitch shift, pitch and chroma, and resolvability jump without the help of time theory. This research contributes to our understanding of auditory perception and has implications for fields such as music theory, cognitive science, and auditory neuroscience.

Jun-ichi TAKAHASHI

8-3-5 Nagaura-ekimae Sodegaura, Chiba, 299-0246, Japan

ORCID iD: 0000-0003-3778-012X

e-mail: takajun@joy.ocn.ne.jp

Keywords: Pitch perception, template matching, missing fundamental, Auditory Scene Analysis, Sound Localization, Sound Integration.

1. Introduction

Music is one of the most mysterious aspects of human cognitive behavior. It consists of two complementary parameters in frequency and time space: pitch and rhythm. Although both are physical continuous quantities, they have a discrete structure in terms of cognition. We focus on pitch perception and examine the neurological origins of the discrete structure. Theories of pitch perception have wavered between place theory and time theory (de Cheveigné, 2004, Oxenham 2013). The theory of pitch perception began with Ohm's resonance hypothesis and place theory (Helmholtz 1954) and was given anatomical evidence by Békésy (1949), who observed frequency-resolved auditory nerves in the cochlea. However, the fact that the fundamental in a complex tone was perceived as a pitch even when its intensity was reduced (Schouten 1940) or masked by noise (Licklider 1954) negated the simple place theory, and a theory focusing on the temporal structure was put forward. On the other hand, low-numbered, resolved harmonics were found to produce a much more robust and salient pitch than high-numbered, unresolved harmonics (Carlyon and Shackleton 1994). This produced another challenge for time theory, which typically did not predict a benefit for low-numbered harmonics over high-numbered harmonics. In the 70s, it was shown that altering the temporal structure of a sound by manipulating the phase of the harmonics that make up the complex tone had no effect on pitch perception (Wightman 1973). The idea of pitch perception as a direct physical measurement was rethought, and in the direction of considering pitch perception as pattern recognition, the pattern-transformation model was proposed by Wightman (1973), the optimum processor theory was proposed by Goldstein (1973), and a template model was proposed by Terhardt (1974). However, it was pointed out that it might be difficult to describe pitch perception by a single mechanism, because analysis by pattern matching also made it difficult to separate harmonics in higher-order complex tones, making template matching difficult (Houtsma and Smurzynski 1989). On the other hand, it was shown that fundamentals were perceived when different higher-order harmonics nf_0 and $(n-1)f_0$ were presented separately to the two ears, indicating that pitch perception is not the result of physical information processing in the auditory peripheral system, but rather in the central system after the crossing of the left and right auditory nerves (Houtsma and Goldstein 1972). Revising the choice between spectral or temporal, an approach emerged that uses a flexible learning process in the neuronal system to calculate the fundamental for acoustic data that describes the two-dimensional information of frequency and temporal information as spectrograms on an equal basis. Shamma and Klein (2000) proposed a network model of pitch perception based on the coupling between harmonics due to the non-linearity response in the cochlear filter. Recently, Saddler Gonzalez, and McDermott. (2021) developed a deep neural network (DNN) model trained to estimate the fundamental frequency and showed that many properties of pitch perception could be quantitatively reproduced by incorporating cochlear filter properties into a DNN. However, Shamma and Klein (2000) noted the shortcomings of the template model. "The lack of convincing biological evidence so far for the existence of these templates or for how they might be generated." The DNN model has a serious problem as it is a teleological procedure to optimize a given parameter. Sound pitch is a very limited piece of acoustic information. How much information about the object can we get from the pitch? Is the cost of creating a pitch perception system worth it? Why does the system choose the pitch as the target parameter for learning and not the center of gravity of the spectrum? Or why does the system perceive discrete harmonics without perceiving the continuous spectral envelope contrary to the Gestalt principle? The main problem with the DNN approach is that today's DNNs are too powerful. If we consider a DNN

as a transformation from an input signal to an output signal, then under the appropriate formulation of the problem and with the appropriate teacher data, the DNN can give a formally accurate solution to almost any problem. The main focus of research using DNNs has shifted from how accurate transformations can be given to what internal mechanisms are responsible for high accuracy, and how they are formed.

How have organisms handled acoustic signals under survival strategies? Auditory perception measures the relative position of the objects and ourselves and provides an impetus for our own behavioral decisions. In insects, signal processing occurs at the first auditory synapse. This computational sparseness does not allow for general-purpose operations, but rather, it limits auditory processing to the selective recognition and localization of significant acoustic signals (Göpfert and Hennig 2016). In humans, on the other hand, sound localization is not just a matter of measuring the relative position of the sound object, but also of extracting the sound object from the sound environment and placing it on a map in the brain as a target for various cognitive operations. This process is known as Auditory Scene Analysis (ASA) (Bregman 1990). In this paper, we return to the mechanism of auditory perception, focus on the harmonic structure of sound objects as the most simple and probable way in neural dynamics to produce such a separation strategy (Trainor 2015), consider a neural network model with non-linear input filters for pitch perception, and consider a harmonic template model. It will be shown that this template operates like the finite state machine (FSM) in the mathematical modeling of computation (Hopcroft, Motwani, and Ullman, 2001) and accounts for fundamental properties in hearing, such as missing fundamental, octave equivalence, pitch shift, pitch and chroma, and resolvability jump at the 10th harmonic, without the aid of a temporal model.

2. Theory

Sound signals are characterized by their amplitude and frequency. Sounds with significantly lower or higher frequencies have no invasive effect on the organism, as resonant interaction with the organism is lost. Each animal has its own species-specific perceptible frequency range. Sounds incident on the ear are converted into vibrations in the lymphatic fluid and stimulate the firing of the corresponding auditory nerves through the vibration of frequency-resolved inner hair cells. On the other hand, signals with excessive amplitude cause irreversible damage to the sensory system, so there is an upper limit to the sensitivity of hearing. Despite this, humans have a large dynamic range of up to 120 dB. It is thought that amplification in the cochlea due to active and mechanical movement in outer hair cells contributes to this (Davis 1983, Dallos 2008, Avan, Büki, and Petit. 2012). In general, non-linear distortions often accompany the amplification with a wide dynamic range. Kadia and Wang (2003) measured the firing of neurons in open marmosets and directly observed second- and third-order nonlinear mode coupling in the sensory response of auditory neurons in the A1 auditory cortex, showing that the central nerve uses mode-transformed harmonics and that there is strong nonlinear mode coupling that cannot be ignored (Wang 2013, Feng and Wang 2017). Aside from the sensory input, it had been known since the end of the 70s that when sound waves are incident on the ear, the ear radiates sound waves with frequencies that are two and three times the frequency of the incident sound wave. This phenomenon is called Distortion Product Otoacoustic Emission (DPOAE) (Kemp 1979, Bian and Chen 2008). The observation of second- and third-order DPOAEs is also direct evidence

of non-linear mode coupling in the auditory organ and shows that the non-linear coupling of the response is sufficiently weak that it is valid to treat intermodal coupling as a perturbation. Although the DPOAE is a mechanical response of the organ to an externally emitted auditory signal and does not directly represent the sensory input, we consider it sufficiently feasible to consider that the properties of the DPOAE are quantitatively correlated with sound perception.

In response to stimulus input from the external world, animals localize objects (in which direction and how distant), judge their attribution (whether to approach, escape from, or ignore), and determine their behavior. Behavior decisions are not mere reflex responses to stimuli but are considered to be the output of a simulation in a model of the external world (representation), according to the location and attribution of the object in the brain through segregation and integration of the input signal. The process of reconstructing the placement of an object in the brain from acoustic signals is called Auditory Scene Analysis (ASA). The localization of objects by sound signals is called sound localization. As sounds of various frequencies arrive at the ear at different times, sound localization is the process of performing ASA by segregating and integrating sounds emanated by objects, using sound intensity, binaural time differences, and other factors as keys.

Based on these previous studies, we propose a model for creating the auditory network through the following processes. In the undifferentiated early stages of hearing, auditory cells are randomly distributed. Signals from multiple auditory cells are sent to neurons to measure their coincidence, which is used to segregate and integrate sounds. Learning in neurons is considered to follow the Hebbian rule. That is, when signal input is superimposed in the presence of teacher signals, the neuron undergoes reinforcement learning. The second- and third-order overtones described above always satisfy the coincidence with the fundamental, so they act as pre-prepared teacher signals in learning. Furthermore, when a sound matching these overtones is superimposed from the outside, the neuron undergoes reinforcement learning for the correlations between the fundamental and these overtones. The reinforcement learning is bidirectional, so for each frequency f, correlations are implemented not only with $2f$ and $3f$, but also with $\frac{1}{2}f$, $\frac{1}{3}f$ and their combinations $2^n3^m f$ over any integer order ($-\infty<n, m<\infty$) (Fig.1).

Since learning is performed in parallel for all frequencies $\{f^{(k)}\}$ consisting of the input sound, where $f^{(k)} = kf_0$ with the fundamental frequency f_0 and harmonic number k=1 to ∞, power series matching templates, $\left[\left\langle 2^n3^m f\right\rangle\right]$ for every frequency f are formed on the distributed neurons. The neurons in the series fire simultaneously for all individual inputs, so frequency discrimination within it is not possible. At the same time, as described below, the third-order nonlinearity $2f_1 - f_2$ adds 5th and 7th matching template harmonics. Finally, harmonic templates, $\left[\left\langle 2^n3^m f\right\rangle, \langle 5f\rangle, \langle 7f\rangle\right]$ are produced. The first column in Table 1 shows the harmonics linked by chains of second- and third-order nonlinearities. Here n=5,7 are prime to 2,3 and each other, so they cannot be directly linked by second- and third-order harmonics with the fundamental. On the other hand, under third-order nonlinearity, a third wave $2f_1 - f_2$ is generated from two different waves of f_1 and f_2. Consider this three-wave pair: when two sounds with f_1 and f_2 are input, a signal with a frequency of $2f_1 - f_2$ acts as a prepared teacher signal. When a sound with $2f_1 - f_2$ is superimposed simultaneously, the neuron undergoes reinforcement learning of the coincidence among the three waves. After learning, when any two of f_1, f_2, and $2f_1 - f_2$ are input, regardless of the order of frequency, signals with the remaining frequency are perceived simultaneously. The blank n=5, 7 in the column of Table 1 can be learned to correlate with the harmonics in the second and third columns by 5=2*4-3, 7=2*8-9, respectively.

Similarly, n=10, 14, and 15 can be learned by the coincidences with the harmonics in the second and third columns by 10=2*9-8, 14=2*16-18, and 15=2*12-9, respectively. Although n=14, 15 can also be decomposed into 2*7, 3*5, the learning efficiency would be lower because $2f_1 - f_2$ learning requires two teacher signals whereas overtone learning requires only one externally provided teacher signal, and the joint strength of the coincidences learned in $2f_1 - f_2$ learning is considered weaker than in overtone learning. Therefore, for n=14 and 15, matching 14=2*16-18 and 15=2*12-9 are considered to be preferred. Also, n=11, 13, 17, and 19 are prime to 2 and 3 each other and cannot be factorized by 2 and 3. It seems that $2f_1 - f_2$ learning is possible using the second- and third-column harmonics as these are decomposed, e.g., 11=2*6-1, 13=2*8-3, 17=2*9-1. However, DPOAE shows a sharp decrease in coupling strength for $f_1/f_2>1.4$ (Bian and Chen 2008). Considering the contribution of amplification by outer hair cells, learning at $f_1/f_2>1.4$ seems to be difficult. Therefore, the harmonics learned by the system should be restricted to the numbers without brackets in the first column of Table 1. Once the learning is established, the system acts as a coincidence detector between the fundamental and the harmonics. The discrete internal structure acquired after the convergence of learning with statistical data will give the system robustness to minor variations in details of the linear properties of input filters, fluctuations in network parameters, etc. We name the harmonic template system derived from the chain of correlation learning Power Series Template (PoST) system. Here, the PoST system includes the 5th and 7th harmonics through the $2f_1 - f_2$ coupling. As mentioned above, learning the $2f_1 - f_2$ coupling is less efficient than direct power series learning. If it is necessary to consider a PoST system without 5th and 7th harmonics, we call the system pre PoST (Fig.1). In the inference process of the PoST system, the input acoustic signal undergoes a step-by-step transformation to a pitch, as will be shown in the next chapter. The production and operation of the PoST system are depicted in Fig.2. The step-by-step transformation can be considered an operation in an FSM if it is regarded as a transition between states. The system acts as a discrete signal processor, and the FSM provides a simple description of operations that is robust for small fluctuations in the parameters.

Up to this point, we have avoided discussing learning through second-order processes. The second-order processes include sum and difference frequencies and could potentially contribute to the learning of 5th, 7th, 11th, and 13th harmonics through relationships using elements of the pre-PoST system, such as 5=3+2, 4+1, 6-1, 7=4+3, 6+1, 8-1, 9-2, and further, 11=9+2, 8+3, 13=9+4. However, when examining the f_1/f_2 dependence of signal strength for CDT (Cubic Difference Tone) and QDT (Quadratic Difference Tone), we observe that CDT shows an increase in signal strength in the range of $f_1/f_2<1.4$, while QDT shows no clear enhancement. Although DPOAE represents the radiation intensity from the inner ear and does not necessarily reflect the received signal strength, it may suggest that in the auditory organ, perception of nonlinear third-order signals under certain special conditions has a stronger preference than that of nonlinear second-order signals. Additionally, if learning through second-order processes were effectively functioning, we cannot deny the possibility of perceiving the 11th and 13th harmonics simultaneously with the 5th and 7th. However, as we will demonstrate later in section 3.3, the perceptual structure of auditory signals aligns well with a group structure generated by 2, 3, 5, and 7, and the perception of 11 and 13 is unnecessary. Based on these two experimental facts, we believe that the perception of the 5th and 7th harmonics would be acquired through learning only third-order processes using pre-PoST elements, while the 11th and 13th harmonics are not sufficiently learned.

In the following, the operation of the PoST system and its characteristics is described in detail.

3. Discussion

3.1. Missing fundamental

It has been a fundamental problem in pitch perception that fundamental missing harmonic sequences induce pitch sensation (Seebeck 1841, Schouten 1940, Licklider 1954).

Examples of the matching scheme of the PoST system are shown in Fig.3. When a sound consisting of multiple harmonics $\{f^{(k)}\}$ is input to a trained neural system, 2^n and 3^m series are evoked separately for each harmonic and are cross-checked for coincidence. (Here, only two tones are shown in Fig.3.) If there is an intersection between the two series, the frequency of the intersection is the fundamental for the input harmonics (Fig.3a). Even when there is no intersection between the respective series of the two input harmonics, the third-order nonlinearity evokes $2f_1 - f_2$ and $2f_2 - f_1$ harmonics from f_1 and f_2, which gives another path for matching in the PoST (Fig.3b). If their networks have an intersection f_0 with the network of f_1 or f_2, the neuronal system returns f_0 as the missing fundamental and f_1 and f_2 are written as the harmonics of f_0. It should be noted that f_0 is unique as 2 and 3 are prime to each other. When the input sound has more than three harmonics, matching is performed in parallel in all pairs of auditory neurons. However, it is not necessary to complete matching on every pair. If a harmonic template generated by the fundamental determined from any two harmonic pairs is consistent with the rest of the harmonic group, no further matching is required, and the sound localization can be terminated. The determination of the pitch signal would give a cue to take the exit action while matching is not completed in other neurons, which would act on the auditory system to interrupt attention to the object and prevent it from wasting further resources for perception. It is noted that the fundamental is not prepared a priori as a generator of the harmonic template (and is sometimes missing in the auditory input), but rather it is more appropriate to think of it as being defined recursively from the input harmonics. The system does not perceive the auditory signal as a whole spectrum, but integrates it into a special single scalar quantity, the pitch of the sound, which allows the target signal to be incorporated into the ASA with minimal resources in perception, just as edge detection in vision does. The parsimony would be critical for quick behavioral decisions.

3.2. Pitch shift

When all of the components in a harmonic complex tone are shifted in frequency by Δf, the perceived pitch of the complex shifts roughly in proportion to Δf. The proportionality factor can be approximated by the reciprocal of the harmonic order of the carrier frequency of the complex tone regarding the pitch fundamental frequency (Schouten 1940, Schouten, Ritsma, and Cardozo. 1962).

The learning of the power series template is carried out in parallel at all frequencies. When a frequency is fixed, the tuning width of its harmonics, learned by power harmonics, cannot be infinitesimally small and always has a finite width. Once the learning is complete, matching at the higher harmonics with a detuning is possible. In listening to complex tones

composed of multiple harmonics of a common fundamental, when the harmonics are shifted by the same frequency, the fundamentals determined by power series template matching for each harmonic will have different values. However, if the modulation is within the tuning width, the brain will prefer to match them with a common frequency. In a trained network, the firing probability of a neuron continuously decreases with greater detuning. Approximating the likelihood for detuning by the simplest additive quadratic function, for example, the likelihood in a three-tone complex is $\sum_{k=n-1}^{n+1}(k\omega_0 + \Delta f - k\left(\omega_0 + \Delta p\right))^2$, and the best estimate is $\Delta p = \Delta f/(n + \frac{2}{3n}) \sim \Delta f/n$ (the first effect of pitch shift), where $k\omega_0 + \Delta f$ is the kth harmonic, including the frequency shift Δf, used in the experiment, and $\omega_0 + \Delta p$ is the virtual fundamental. It is known that the proportionality factor obtained experimentally deviates systematically from $1/n$ (the second effect of pitch shift) (Schouten, Ritsma, and Cardozo. 1962, Smoorenberg 1970). Its origin could be attributed to the characteristics of the likelihood function.

3.3. Pitch and Chroma

It is well known that sounds with twice the frequency are perceived as having the same pitch. (Octave circularity in pitch perception) (Deutsch 2010, Shepard 1982).

In our model, harmonics that differ by a factor of 2^n in frequency are integrated into a single series during learning and are no longer distinguishable. This is consistent with octave circularity. Similarly, harmonics with different 3^m fold frequencies are integrated into a single series, which is known as the perfect 5th consonance. Shepard geometrically represented the periodicity, including octaves and perfect 5ths, in pitch with the double helix model (Shepard 1982). In general, the harmonics of the sounds that make up the object sound consist of harmonics of the natural numbers of orders of the fundamental harmonic. On the other hand, the power series templates contain harmonics with negative integer powers, and the perception of subharmonics would be learned simultaneously due to the bidirectionality of the learning, whereas they do not contribute to sound localization. As noted already, 5 and 7 also work as generators of the tone system of pitch perception. The subharmonics can be folded back to a region above the fundamental frequency by multiplying by the powers of 2, 3, 5, or 7. These are frequencies with a rational ratio to the fundamental and define a chroma. For example, multiplication of 3 and folding into an octave produces the Pythagorean scale. Multiplication of 3 and 5 and folding into an octave produces Just Intonation. Which chroma is preferred or survives will depend on culture, and the resulting musical scales.

3.4. Resolvability

The fundamental discrimination threshold (F0DT) of the complex harmonic tone, which consists of several harmonics, shows a critical threshold around N=10 (Houtsma and Smurzynksi 1990, Bernstein and Oxemham 2003). Bernstein and Oxemham investigated it using amplitude-modulated tones consisting of three consecutive harmonics, showing high discrimination performance at N≦10, with a sharp drop at N>10. However, there is no significant change in discrimination performance at much higher orders. Because the components of a harmonic complex are equally spaced on a linear frequency scale, but the absolute bandwidths of auditory filters increase with increasing center frequency, the density of

harmonics per auditory filter increases with increasing harmonic number. As a result, low-order harmonics are resolved from one another, but higher-order harmonics begin to interact within single auditory filters and eventually become unresolved. In this case, however, resolvability will gradually decrease, and the jump at the 10th order will not appear. Then, it was proposed that pitch perception was attributed to matching with the harmonic template for low-order harmonics and reading the period pattern of the acoustic waveform for higher-order harmonics.

This discontinuous change has been one of the biggest problems in pitch perception theory. Let us look at this problem from the standpoint of power series template matching. Consider a complex tone consisting of several harmonics (Table 2). If the tone contains two consecutive orders of harmonics, the $2f_1 - f_2$ coupling produces the perception of two harmonics below and above the orders at the same time (underlined). A power series template matching is performed among a total of four harmonics. In each complex tone, the 2^n and 3^m power series template intersections are calculated consistently by the harmonics in bold, which give matching solutions listed in the third column. The matching calculation has a solution for all combinations up to N≦10. In this case, it is not necessarily the fundamental that is determined by the calculation, but the doubling for the pairs (5,6) and (6,7) and the tripling for the pairs (10,11). The multiples do not match the original harmonic pairs. They are folded again into the fundamental to complete the template matching.

Even if the complex tone contains more harmonics, if they match a harmonic template, they will be integrated into the same power series templates, giving the same fundamentals. The matching calculation using two harmonics except (16,17) and (17,18) does not have a solution for N≧11, but by using three consecutive harmonics, it is possible to have solutions up to (18,19,20). An increase in the number of harmonics for the matching calculation decreases the efficiency of the matching calculation, and therefore the fundamental discrimination thresholds for N≧11 would have to be larger than that for N≦10. Our answer to the problem of resolved and unresolved harmonics for pitch perception is the special structure of the matching template itself rather than the transition to the perception of time structure.

3.5. Music, Language, and PoST system

We have investigated the embryological origin of acoustic signal perception and derived the structure of the perceptual system generated by the perturbative nonlinearity associated with the active amplification of small acoustic signals and reinforcement learning in the auditory neural system. Our model has three features.

1. It uses frequency coding and has no time coding.
2. It can perceive not only harmonics but also subharmonics. The perception of the fundamental frequency as the identity element and its harmonics and inverse elements brings a group-theoretical structure to pitch perception with the generating elements of 2, 3, 5, 7.
3. The harmonic template has deficiencies for orders higher than 10, and a jump of F0DT occurs at the 10th order when pitch inference is performed.

The elements of the group can be written as $2^n3^m5^p7^q f_0$, where m, n, p, q are integers ($-\infty < m, n, p, q < \infty$) and f_0 is the fundamental frequency. Our auditory organ has a finite bandwidth and resolution, so the sound structure we can perceive

is limited to a small subset of the above group. This subset will correspond to the number-theoretical structure seen in the harmony of music. The octave circularity is the equivalence class of integers modulo 2, and the undertone perception that has been excluded in Western music theory is the inverse element of the group. When group elements belonging to the same f_0 are presented simultaneously within the auditory relaxation time, the brain perceives strong correlations between the sounds. Although our model assumes a single source and temporal coincidence, these constraints can be relaxed because the memory of auditory perception has a finite relaxation time. Relaxation of the single sound source may lead to chords, while relaxation of the temporal coincidence may lead to scales.

On the other hand, it is well accepted that hearing exploits spatiotemporal coding (Shamma and Klein, 2000, Cedolin, and Delgutte. 2010, Saddler et.al. 2022). Our perception of music clearly demonstrates the auditory system's employment of both spatial and temporal coding mechanisms. However, a critical question arises: in the context of the evolution of pitch perception, was it necessary for frequency and temporal coding functions to coexist, despite the potential for strategies to enhance auditory filter resolution? Furthermore, is the intersection of frequency and temporal coding at the 10th order a necessity or merely a coincidence? Why is there pitch perception in the first place? In our framework, pitch perception represents merely a part of the overall auditory function. We propose that the primary biological adaptation phenomenon is template matching aimed at sound localization in ASA. Pitch perception, in this context, emerges as a byproduct of this template matching process, having no adaptation target.

In our model, temporal coding is not necessary to explain the fundamental characteristics of pitch perception, particularly its emergence. However, due to the finite relaxation time in acoustic signal perception, pitch correlations are also provided in the temporal domain. The temporal variations in pitch (and its complementary timbre) introduce new degrees of freedom, allowing for diverse contexts in acoustic signals. A significant distinction between auditory and visual perception lies in the former's capacity to be coupled with active cognitive functions such as vocalization, whereas the latter is predominantly limited to passive cognitive processes. This coupling of auditory perception with active vocal production represents a fundamental difference in the cognitive architecture of these sensory modalities. Although pitch does not have a direct adaptive target, it can find a new adaptive target: communication. Temporally structured sound signals will evolve into music and language.

In recent years, the relationship between musical scale structures and vocalization in language has attracted significant attention. Brown and Phillips have proposed a theory suggesting that the physiological structure of the larynx restricts pitch intervals, resulting in musical scales composed of a finite number of tones (Brown and Phillips 2023). Schwartz et al. extracted thousands of voiced segments from a speech database, determined their spectra, and demonstrated a preference for pitches corresponding to musical scales in speech (Schwartz, Howe, and Purves 2003). Bowling and Purves attributed the origin of musical scales to vocalization (Bowling and Purves 2015). Nevertheless, it should be recognized that the commonalities between the sound structures of music and language do not necessarily indicate a causal link. According to our model, the perceptual structure of auditory signals arises biologically for sound localization in ASA, generating harmonic templates from the group-theoretic structure having generators of prime numbers 2, 3, 5, and 7. Concurrently, pitch perception is acquired as a byproduct having no adaptation. Subsequently, pitch perception, which initially lacked an adaptive target, evolves to find an adaptive target of culture (in our context, inter-individual

communication), leading to the development of musical scales and other sound structures, ultimately evolving into music and language. We propose the hypothesis that music and language are exaptations of sound localization in ASA. This perspective offers a novel approach to understanding the evolution of complex auditory phenomena and their cultural significance.

4. Conclusion

We have approached the problem of pitch perception with the help of nonlinear dynamics of the cochlear amplifier, evolution of hearing, auditory psychology based on ASA, and discrete mathematics. We started from the perturbative second- and third-order nonlinearity of cochlear amplification of small auditory signals and considered the chains of reinforcement learning of the neural network, which produced a power series harmonic template. The third-order nonlinearity, $2f_1 - f_2$, coupling filled some blanks in the raw power series harmonic template, which gave the harmonic template system named PoST. The chains enabled the perception of coincidence between the fundamental and higher harmonics at the cost of losing octave perception, resulting in octave equivalence. The $2f_1 - f_2$ coupling filled the 5th and 7th blanks of the raw power series harmonic template, resulting in an improvement of F0DT for complex tones below the 10th. We attributed the characteristic features of pitch perception to the constraints on the ideal DNN. Our model could give not only explanations about the characteristic features of pitch perception without the help of time theory but also the answers to fundamental problems of pitch perception. What is pitch perception? The success of sound integration of emanated sound from a single object. Why was pitch perception born? Reinforcement learning of sound localization using second- and third-order nonlinearity as teacher signals. Why is pitch classified into discrete chroma? Subharmonic perception from the bidirectional coupling in harmonic folding.

Figures and Tables

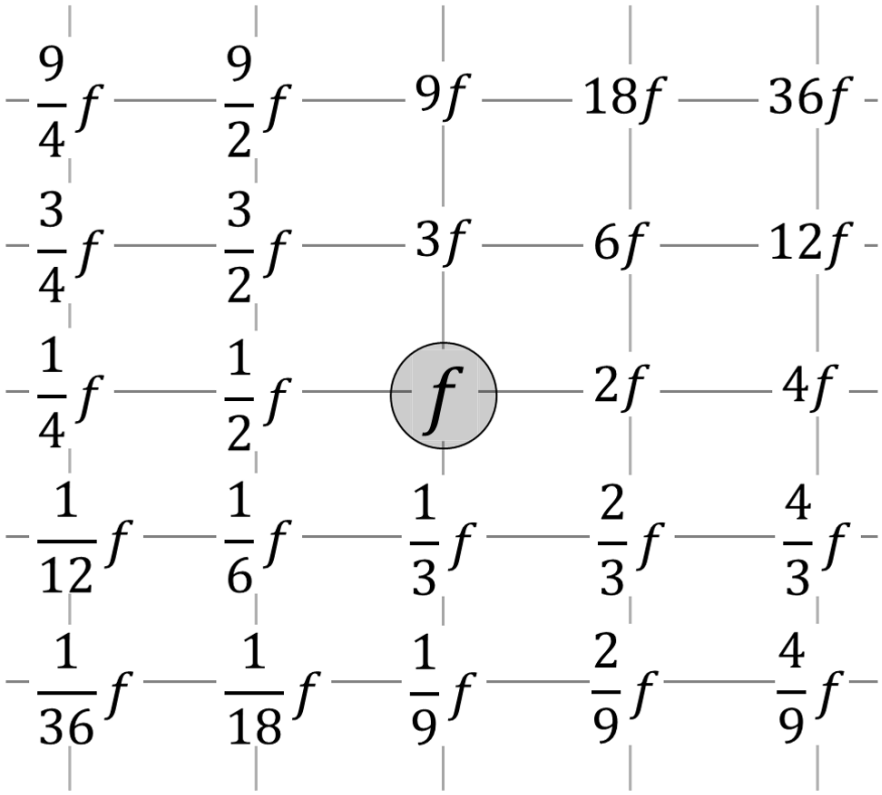

Figure 1. Network of evoked signals: When a sound with frequency f is heard, neurons associated with fire simultaneously. Under the third-order nonlinearity of $2f_1 - f_2$, $5f$ and $7f$ neurons can be fired by the recursive interaction from $2^n3^m f$ neurons (orthogonal to 2^n3^m plane). If a sound with $2^n3^m f$ is overlapped, the coincidence with f is perceived.

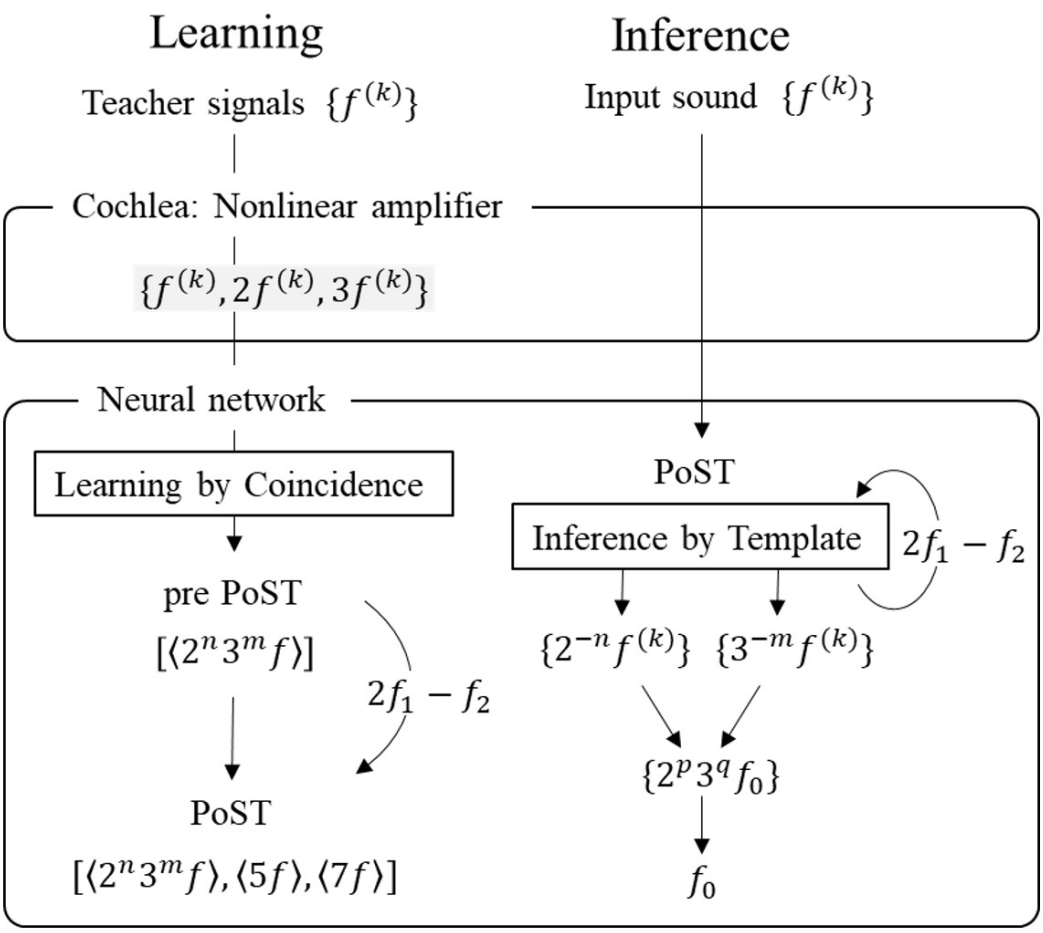

Figure 2. Learning and inference of PoST system: $\{f^{(k)}\}$ are harmonic series of kf_0 with a fundamental frequency f_0 and k=1 to ∞. $\{f^{(k)}, 2f^{(k)}, 3f^{(k)}\}$ are combined series with second- and third-order harmonics. $\left[\left\langle 2^n 3^m f\right\rangle\right]$ are power series harmonic templates for each f. $\left[\left\langle 2^n 3^m f\right\rangle, \langle 5f\rangle, \langle 7f\rangle\right]$ are the PoST templates to which 5th and 7th matching harmonics are added. In the learning process, nonlinear cochlear amplification provides second and third harmonics, and they are exploited by the NN as teacher signals, producing PoST. In the inference process, PoST outputs chains of the power series of second- and third-order couplings from each $f^{(k)}$. If they have an overlap, they are integrated into a single series with the fundamental frequency f_0. If f_0 is consistent with all of $f^{(k)}$, f_0 is output as the fundamental frequency of the input sound.

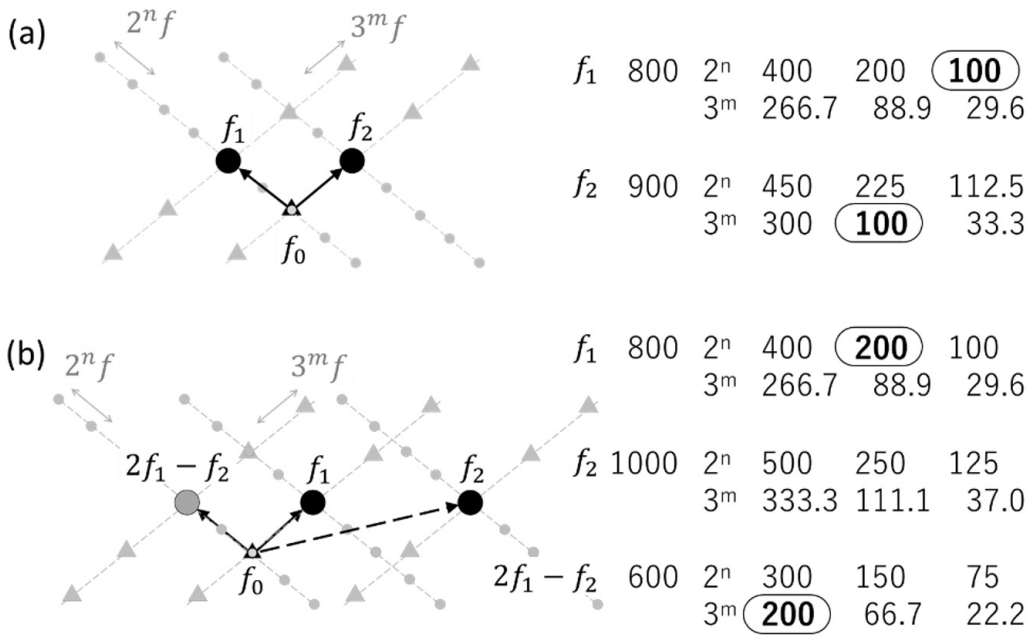

Figure 3. Matching scheme of PoST system: (a) when 800 Hz and 900 Hz tones are input simultaneously, each evokes 2^n and 3^m power series templates, respectively. The intersection 100 Hz (bold) is uniquely determined from the former 2^n template and the latter 3^m template. (b) When the 800 Hz and 1000 Hz tones are input simultaneously, there is no intersection in the power templates evoked from each. However, they evoke a 600 Hz tone by the $2f_1 - f_2$ coupling, defining the intersection 200 Hz (bold) uniquely from the 2^n template of 800 Hz and the 3^m template of 600 Hz.

Table.1 Harmonic template learned by the second- and third-order nonlinearities: First column: the order of harmonics, second and third columns: the number of iterations of second-order and third-order nonlinear coupling required to access the harmonics, fourth column: the combinations of those defined from the second and third columns to access the harmonics by the $2f_1 - f_2$ coupling. Here, those of f_1/f_2>1.4 are not written (see text). The numbers in brackets are not accessible by the second- and third-order nonlinearities.

N	2^n	3^m	$2f_1 - f_2$		N	2^n	3^m	$2f_1 - f_2$
1					14			2*16-18
2	1				15			2*12-9
3		1			16	4		
4	2				(17)			
5			2*4-3, 2*2+1		18	1	2	
6	1	1			(19)			
7			2*8-9, 2*3+1		20			2*18-16, 2*16-12
8	3				(21)			
9		2			(22)			
10			2*9-8, 2*8-6		(23)			
(11)			(2*4+3)		24	3	1	
12	2	1			…			
(13)								

Table 2. Calculation of fundamentals in complex tones with consecutive harmonics: Left: two tone complexes (N≦11), right: three tone complexes (N≧10). First column: given harmonic combinations. Second column: evoked harmonics by the $2f_1 - f_2$ coupling (underlined) and those used for 2^n and 3^m template matchings (bold). Third column: harmonics determined as the intersection of power series template matching. The determined harmonics are folded into fundamental under octave and perfect 5th circularity to give a pitch perception so that the harmonic template is consistent with the harmonic series in the first column. Up to (10,11), only two harmonics are enough to determine the fundamental, whereas three harmonics are necessary above that, up to (18,19,20).

Given	Evoked	Matched		Given	Evoked	Matched
(2,3)	(**2**,**3**,**4**)	1		(10,11,12)	(**8**,**9**,10,11,**12**,13,14)	1
(3,4)	(**2**,**3**,**4**,5)	1		(11,12,13)	(**9**,10,11,**12**,13,14,15)	3
(4,5)	(**3**,**4**,5,**6**)	1		(12,13,14)	(10,11,**12**,13,14,15,**16**)	4
(5,6)	(**4**,5,**6**,7)	2		(13,14,15)	(11,**12**,13,14,15,**16**,17)	4
(6,7)	(5,**6**,7,**8**)	2		(14,15,16)	(**12**,13,14,15,**16**,17,**18**)	2
(7,8)	(**6**,7,**8**,**9**)	1		(15,16,17)	(13,14,15,**16**,17,**18**,19)	2
(8,9)	(7,**8**,**9**,10)	1		(16,17,18)	(14,15,**16**,17,**18**,19,20)	2
(9,10)	(**8**,**9**,10,11)	1		(17,18,19)	(15,**16**,17,**18**,19,20,21)	2
(10,11)	(**9**,10,11,**12**)	3		(18,19,20)	(**16**,17,**18**,19,20,21,22)	2
(11,12)	(10,11,**12**,13)	×		(19,20,21)	(17,**18**,19,20,21,22,23)	×
…				…		

Statements and Declarations

Conflict of Interest

The author and the present work have no intellectual or financial support from anyone and no conflict of interest with any other works.

Open Practices Statement

Neither of the studies reported in this article has been pre-registered. No data other than those presented here are available.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author used none of generative AI and AI-assisted technologies other than language editing services (DeepL, Grammarly).

References

- Avan, P., Büki, B., and Petit, C., (2013) “Auditory Distortions: Origins and Functions,” Physiol. Rev.,*93*, 1563–1619. https://doi.org/10.1152/physrev.00029.2012
- Békésy, G., (1949) “The vibration of the cochlear partition in anatomical preparations and in models of the inner ear,” J. Acoust. Soc. Am. *21*, 233-245. https://doi.org/10.1121/1.1906502
- Bernstein, J.G., and Oxenham, A. J., (2003) “Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number?” J. Acoust. Soc. Am. *113*(6), 3323-3334. https://doi.org/10.1121/1.1572146
- Bian, L., and Chen, S., (2008) “Comparing the optimal signal conditions for recording cubic and quadratic distortion product otoacoustic emissions,” J Acoust Soc Am., *124*(6),3739–3750. https://doi.org/10.1121/1.3001706
- Bowling, D. L., and Purves, D., (2015). “A biological rationale for musical consonance,” Proceedings of the National Academy of Sciences, *112*(36), 11155-11160. https://doi.org/10.1073/pnas.1505768112
- Bregman, A. S., (1990) “Auditory scene analysis: The Perceptual Organization of Sound. Cambridge,” MA: MIT Press. ISBN 9780262022972.
- Brown, S., and Phillips, E., (2023). “The vocal origin of musical scales: the Interval Spacing model,” Frontiers in Psychology, 14, 1261218. https://doi.org/10.3389/fpsyg.2023.1261218
- Carlyon, R. P., and Shackleton, T. M., (1994) “Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms?” J. Acoust. Soc. Am., *95*, 3541-3554. https://doi.org/10.1121/1.409971
- de Cheveigné, A., (2004) “Pitch perception models-a historical review,” CNRS-Ircam, Paris, France.
- Cedolin, L., and Delgutte, B. (2010) “Spatiotemporal representation of the pitch of harmonic complex tones in the auditory nerve,” Journal of Neuroscience, *30*(38), 12712-12724. https://doi: 10.1523/JNEUROSCI.6365-09.2010

- Dallos, P., (2008) “Cochlear amplification, outer hair cells and prestin,” Curr. Opin. Neurobiol.,*18*(4), 370–376. doi:10.1016/j.conb.2008.08.016.
- Davis, H., (1983) “An active process in cochlear mechanics,” Hearing Research,*9*(1), 79-90. https://doi.org/10.1016/0378-5955(83)90136-3
- Deutsch, D., (2010) “The Paradox of Pitch Circularity,” Acoustics Today, July, 8-15
- Feng, L., and Wang, X., (2017) “Harmonic template neurons in primate auditory cortex underlying complex sound processing,” Proc. Natl. Acad. Sci. USA, *114*(5), E840-E848, https://doi.org/10.1073/pnas.1607519114
- Goldstein, J. L., (1973) “An optimum processor theory for the central formation of the pitch of complex tones,” J. Acoust. Soc. Am., *54*, 1496-1516. https://doi.org/10.1121/1.1914448
- Helmholtz, H., (1954) “On the Sensations of Tone: As a Physiological Basis for the Theory of Music,” New York: Dover.
- Hopcroft, J. E., Motwani, R., and Ullman, J. D. (2001). Introduction to automata theory, languages, and computation. Acm Sigact News, *32*(1), 60-65.
- Houtsma, A. J. M., and Goldstein, J. L., (1972) “The Central Origin of the Pitch of Complex Tones: Evidence from Musical Interval Recognition,” J. Acoust. Soc. Am., *51*, 520-529. https://doi.org/10.1121/1.1912873
- Houtsma, A. J. M. and Smurzynski, J., (1989) “Pitch of complex tones with many high-order harmonics,” J. Acoust. Soc. Am., *85*, S142. https://doi.org/10.1121/1.2026776
- Houtsma, A. J. M. and Smurzynski, J., (1990) “Pitch identification and discrimination for complex tones with many harmonics,” J. Acoust. Soc. Am., *87*, 304-310. https://doi.org/10.1121/1.399298
- Kadia, S. C., and Wang, X., (2003) “Spectral Integration in A1 of Awake Primates: Neurons With Single- and Multipeaked Tuning Characteristics” J. Neurophysiol. *89*, 1603–1622. https://doi.org/10.1152/jn.00271.2001
- Kemp, D. T., (1979) “Evidence of mechanical nonlinearity and frequency selective wave amplification in the cochlea,” Arch. Otol. Rhino. Laryngol. *224*, 37–45. https://doi.org/10.1007/BF00455222
- Licklider, J. C. R., (1954) “Periodicity pitch and place pitch,” J. Acoust. Soc. Am.,*26*, 945. https://doi.org/10.1121/1.1928005
- Oxenham, J. A., (2013) “Revisiting place and temporal theories of pitch,” Acoust. Sci. & Tech.*34*, 6. https://doi.org/10.1250/ast.34.388
- Saddler, M. R., Gonzalez, R., and McDermott, J. H., (2021) “Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception,” Nat Commun *12*, 7278. https://doi.org/10.1038/s41467-021-27366-6
- Schouten, J. F., (1940) “The Perception of Pitch,” Phillips Tech. Rev,*5*, 286-294.
- Schouten, J. F., Ritsma, R. J., and Cardozo, B. L., (1962) “Pitch of the Residue,” J. Acoust. Soc. Am*34*, 1418-1424. https://doi.org/10.1121/1.1918360
- Schwartz, D. A., Howe, C. Q., and Purves, D., (2003). “The statistical structure of human speech sounds predicts musical universals,” Journal of Neuroscience, *23*(18), 7160-7168. https://doi.org/10.1523/JNEUROSCI.23-18-07160.2003
- Seebeck, A. (1841) “Beobachtungen uber eineige bedingungen der entstehung von tonen,” Ann. Phys. Chem.*53*, 417-436,
- Shackleton, T. M., and Carlyon, R. P., (1994) “The role of resolved and unresolved harmonics in pitch perception and

frequency modulation discrimination," J. Acoust. Soc. Am. *95*(6), 3529-3540. https://doi.org/10.1121/1.409970

- Shamma, S., and Klein, D., (2000) "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system," J. Acoust. Soc. Am., *107*, 2631-2644. https://doi.org/10.1121/1.428649
- Shepard, R. N., (1982) "Geometrical Approximations to the Structure of Musical Pitch," Psychological review*89*(4), 305-333. https://doi.org/10.1037/0033-295X.89.4.305
- Smoorenburg, G. F., (1970) "Pitch Perception of Two-Frequency Stimuli," J. Acoust. Soc. Am.*48*, 924-942. https://doi.org/10.1121/1.1912232
- Terhardt, E., (1974) "Pitch, consonance, and harmony," J, Acoust, Soc. Am.*55*, 1061-1069. https://doi.org/10.1121/1.1914648
- Trainor, L. J., (2015). "The origins of music in auditory scene analysis and the roles of evolution and culture in musical creation." Philosophical Transactions of the Royal Society B: Biological Sciences, *370*(1664), 20140089. https://doi.org/10.1098/rstb.2014.0089
- Wang, X., (2013) "The harmonic organization of auditory cortex," Front. Syst. Neurosci. 7:114. https://doi.org/10.3389/fnsys.2013.00114
- Wightman, F. L., (1973) "The pattern-transformation model of pitch," J. Acoust. Soc. Am*54*, 407-416. https://doi.org/10.1121/1.1913592