A Better Theory of Music, for Physicists

Word Count: 2998

Introduction

The term "music theory" would suggest a body of knowledge which can be used to explain musical phenomena[1]. Let us examine this hypothesis by considering a passage taken from *The AB Guide to Music Theory* [1]:

"*Virtually all pieces of music written before the early 20th century do not use all the black and white notes but only a selection of them … the most common scale of all can be found by playing just the white notes on the keyboard …*" (p. 11)

There are various problems with this to us physicists. Why do we play just the white notes? Why is the major scale so common? Why are the black and white notes a *complete* collection of notes? The book, and virtually all music theory textbooks, does not seek to answer these kinds of questions for us.

Unfortunately, music theory is not an adequate scientific theory of anything. Rather, it is a set of tools that musicians use to write and deconstruct "nice" sounding music with. How the ancient musicians came up with these—the basis of music theory—is deeply connected to acoustics, the branch of physics dealing with sound. There clearly is a need for some better theory of music which *derives* these sets of tools from first principles. This is what we will do here: we will explore how to obtain the basic Western scales, chords, and tuning systems, and through this we will seek to answer the questions posed. Furthermore, we will explore the cause and implications of a variety of musical phenomena: for example, sound timbre, the feelings associated with major/minor chords, or the impracticality of piano tuning.

The Nature of Pitched Sound

We begin by considering the most basic device in music: a taut string. The equation of motion for the displacement ψ of a string of length L at the origin, is the one-dimension wave equation, with speed $c = \sqrt{T/\rho}$, where T is the tension and ρ the linear density. Its solutions, given the boundary conditions that the ends must be fixed, can be obtained by separation of variables as [2]:

$$\psi(x,t) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right) \sin\left(\frac{n\pi ct}{L} + \phi_n\right) ,$$

where the A_n and ϕ_n are constants which depend completely on the initial conditions (how the string is plucked) and may be obtained by Fourier analysis.

The important thing to note, from these solutions, is that the string vibrates with a linear sum of *discrete* frequencies $f_n = \frac{n\sqrt{T/\rho}}{2L}$. The sinusoidal waves with these frequencies are known as the nth harmonics; the first harmonic is given a special name: the fundamental. The infinite sum of all such harmonics is known as the harmonic series[2]. Vibration of the string at these frequencies cause disturbances in ambient air pressure near the string, which evolves in time matching the string vibration. These air pressure variations are perceived as sound, which is distinctively *pitched* [3].

[1] Or, if you're a classical musician, music theory would probably correlate to the few months of forced dreary studying that eventually leads to an exam that one must pass, to progress onto Grade 6 and beyond on their instrument. I think it's safe to say that it is mostly due to this disturbing experience, that music theory is unanimously hated amongst amateur musicians.

[2] Actually, *the* harmonic series is defined as $\sum_n 1/n$, which is the sum of all normalised discrete *wavelengths* λ_n. We will not use this definition here as sound waves are almost always characterised by frequencies, rather than by wavelengths.

Instruments that use strings to produce sound are called chordophones. It can be further shown [4] that for aerophones, instruments that produce sound by a vibrating column of air in a tube, open–open boundary conditions yield the same harmonic series solutions, whilst closed–open boundary conditions yield $f_n = \frac{(2n-1)c}{4L}$, which can be interpreted as a harmonic series with missing even harmonics. It turns out that most common Western pitched instruments are chordophones or aerophones. Furthermore, pitched acoustic instruments that produce sound which does not contain harmonics are very rare. An example is whistling, which produces sound that is approximately sinusoidal [5].

It is exactly the distribution of the amplitudes of the harmonics that describes the **timbre** of sound, which explains why instruments sound different, even when playing the same note. Timbre may change across an instrument with pitch or amplitude of the fundamental, typically due to nonlinear behaviour of the sound production system [4, 6]. The piano is relatively uniform in timbre with pitch across a given short range of frequencies, but not with loudness. This explains why transposition, the act of shifting all notes' frequencies by a (small) constant ratio in a piece of music, works particularly well on the piano. However, playing "very loudly" will make a distinctively harsh sound, which can still be heard even if the music is recorded and its amplitude reduced.

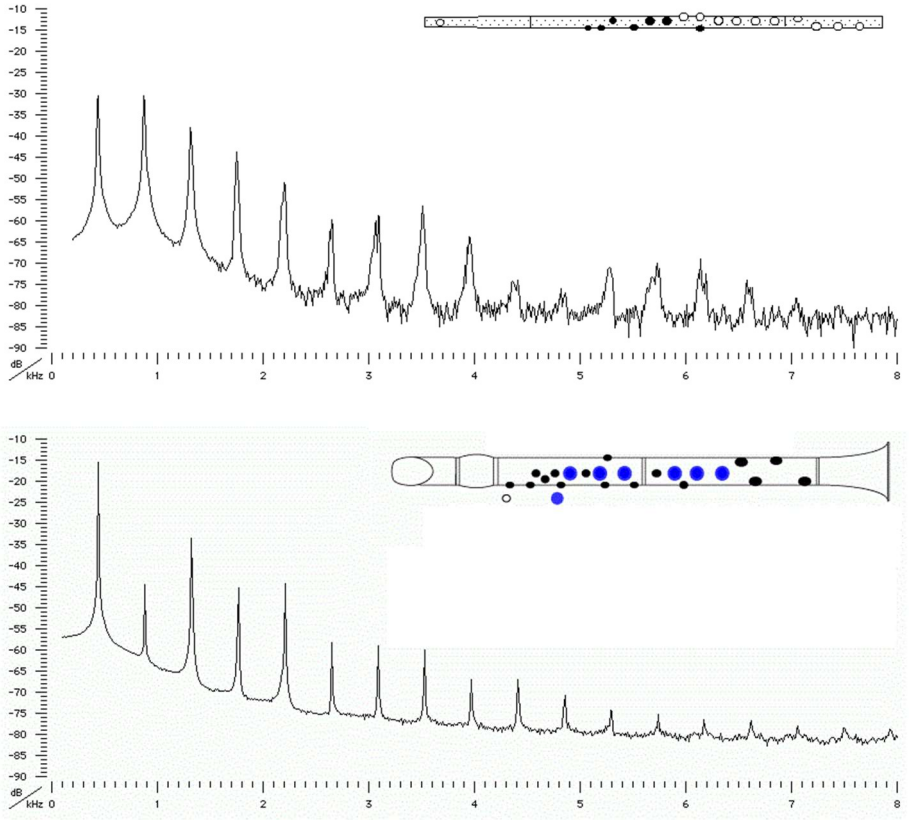

Figure 1: the frequency spectra of a flute, and a clarinet, respectively, playing $A_4 = 440$ Hz [11, 12]. The flute's frequency spectrum contains relatively rounded peaks at the harmonics, which decays exponentially after the second, whilst the clarinet, an instrument well-approximated as a closed–open tube, contains sharp peaks that most notably lack the second harmonic.

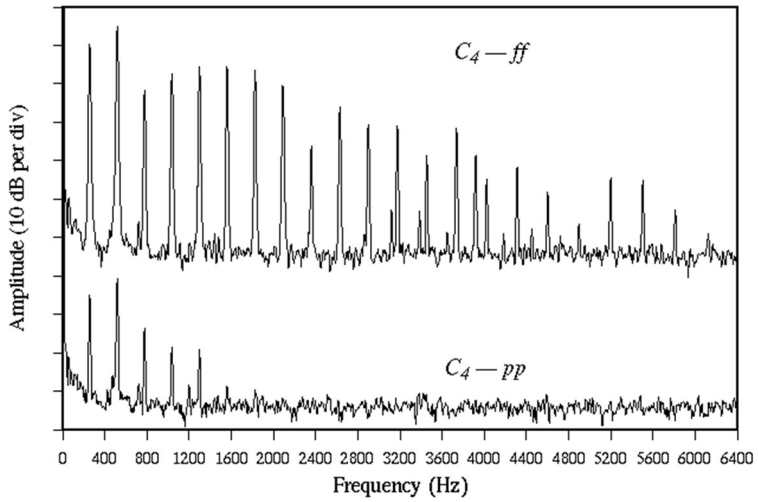

Figure 2: the frequency spectra of a piano playing $C_4 \approx 262$ Hz, firstly at *ff* (*fortissimo*: "very loudly"), then at *pp* (*pianissimo*: "very quietly") [6].

Consonance and Dissonance

Before we go on to derive basic melody and harmony, we need to introduce the acoustical phenomenon of **beating**. Consider the sum of two sinusoidal sound waves with the same amplitude A, and *slightly different* angular frequencies $\omega_1 = 2\pi f_1$ and $\omega_2 = 2\pi f_2$. It can be shown [2] that

$$A \sin \omega_1 t + A \sin \omega_2 t = 2A \sin \left(\frac{\omega_1 + \omega_2}{2} t \right) \cos \left(\frac{\omega_1 - \omega_2}{2} t \right) .$$

The effect is therefore hearing sound at the frequency average $(f_1 + f_2)/2$ which varies in amplitude at the frequency difference $(f_1 - f_2)$. If the waves had different amplitudes, it can be shown [7] that an amplitude envelope corresponding to the beat frequency $(f_1 - f_2)$ remains present. The resulting sound, with its modulating amplitude, is not particularly pleasant. Roughly speaking, beats become unnoticeable if $(f_1 - f_2) > 20$ Hz [4].

Beating forms part of the theory of dissonance. The other part of the theory lies in **roughness** felt between notes of similar frequencies [4]. Roughness is a different phenomenon from beating and can simultaneously occur with beating. The part of our ear which allows us to distinguish frequencies is the basilar membrane, which vibrates at different locations dependent on the frequency of incoming sound. The points on the basilar membrane can be thought of as bandpass filters with a characteristic width, called the critical bandwidth. Experiments show that a good approximation for the critical bandwidth is slightly less than 20% of the centre frequency. If two notes were simultaneously played with a frequency difference such that their critical bandwidths *overlapped*, then the vibration at the region of overlap in the basilar membrane, which can be thought of as unwanted noise, becomes reinforced. We feel this as a sensation of roughness.

The Octave, the Major Triad and the Major Scale

Armed with this knowledge, let us consider the question: given a single note with a specific frequency, how do we select other notes to use? This will allow us to construct a collection of notes, so we can write harmony (notes sounding together) and melody (notes sounding successively in time).

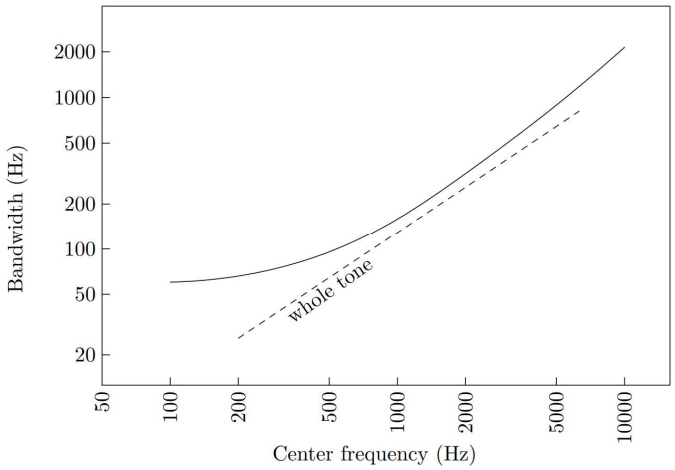

Figure 3: a sketch of the bandwidth of the basilar membrane as a function of centre frequency [4]. The dashed line represents a whole tone/major second, which is 12.5% of the centre frequency.

Let us call the given note C. To avoid dissonance and have consonance (which we define as the lack of dissonance), between C and another note when played together, we must consider the effect of beating and roughness between C, the other note, and their harmonics. We know this because we showed that pitched sound, in general, contains many harmonics. So, which notes are completely consonant with C? Well, any note which has double the frequency, or any integer multiple thereof, satisfies the criteria. The doubling of frequency is known as an **octave** (up). The criteria are satisfied, because all harmonics of the octave line up exactly with the even harmonics of C.

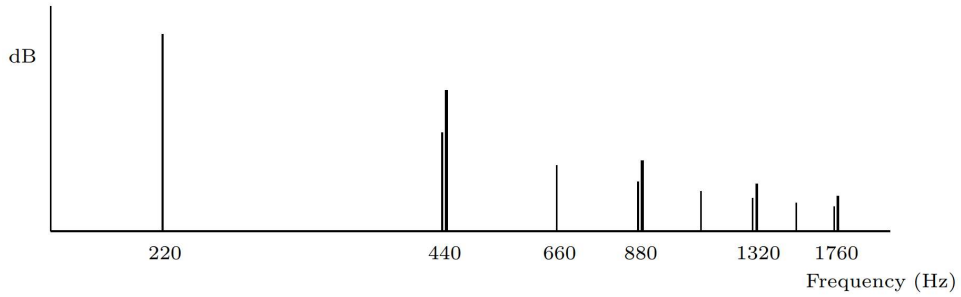

Figure 4: a frequency spectrum of two notes of frequencies 220 Hz and 445 Hz [4], which are related by slightly more than an octave. The harmonics of the two notes are clearly dissonant.

It is precisely because of this, that the octave is interpreted by the brain as really being the exact same note as C, but higher. When we're asked to sing back some melody that's perhaps too low for us, what we're really doing is singing the same melody, transposed an octave up. This phenomenon is known as octave equivalence [4]. The human hearing range is approximately 20 Hz to 20 kHz, implying we can obtain at least nine full octaves for use[3]. ISO defines A_4 [8] as any note with the fundamental at exactly 440 Hz (for historical reasons), where the subscript n indicates the nth octave. The note A_3 would correspond to an octave down, or 220 Hz, for example.

[3] The human sight range, in contrast, is far more restricted, being just short of one octave at approximately 405 to 790 THz [4]. It may be tempting to use this to explain why red seems to join up with violet, but since light doesn't really contain harmonics, it implies this explanation is incorrect.

So, we have nine notes to work with. Can we do better than that? By the octave equivalence principle, we can normalise notes we want to further obtain, by multiplying or dividing their frequencies by integer powers of two, such that they lie within the range of the octave. To go further, we would have to insist on having some dissonance due to roughness. Which notes are the least dissonant? Let us consider the two closest higher harmonics which do not repeat by octaves [4, 9]. The third harmonic has a normalised frequency ratio to the fundamental of 3/2. This frequency ratio is known as the perfect fifth, and the perfect fifth up from C is G. The fifth harmonic similarly gives a frequency ratio of 5/4, which is the major third, or E. We stop here and do not consider the seventh harmonic or above because we know the harmonics' amplitudes decay, sometimes very rapidly on certain instruments, so these do not contribute significantly to dissonance [9].

Now consider playing C, E, and G simultaneously to form a chord. This specific combination of frequency ratios (4:5:6) is known as the **major triad**, "I". The sound is very pleasant, due to fact the three notes are the lowest parts of the same harmonic series: their harmonics do not produce strong beats and cause minimal roughness. We can then build two further major triads based on different notes: one which has the bottom note as G ("V": G–B–D), and one which has the top note as C ("IV": F–A–C) [4], such that we can repeatedly build chords symmetrically in both directions, should we want to. This means the chord I remains at the centre. Thereby we obtain seven notes in an octave, not including the octave itself. This collection of notes is known as the **major scale**. The white keys of a (justly tuned) piano keyboard correspond to exactly this scale.

Note	C	D	E	F	G	A	B	C_{+1}
Frequency Ratio	1	9/8	5/4	4/3	3/2	5/3	15/8	2
Equivalent Decimal	1.000	1.125	1.250	1.333…	1.500	1.666…	1.875	2.000

Table 1: Frequency ratios between notes of the (just) major scale [4].

The Chromatic Scale

Taking the base two logarithm (as the octave correspond to a multiplication by two) of the frequency ratios in the major scale reveals that there are five large gaps and two small gaps between adjacent notes [9]. We can fill in these large gaps with five notes of appropriate frequency ratios. There are various ways this can be done [4], but the ratios, like notes from the major scale, must be of small integers. The black keys of a (justly tuned) piano keyboard would correspond to these five extra notes.

```
C       D       E   F       G       A       B   C
+----+----+----+----+----+----+----+----+----+----+
0    1    2    3    4    5    6    7    8    9    0
```

Figure 5: a plot of $\log_2$ of the frequency ratios in the major scale, rounded to the nearest 0.02. [9].

These twelve notes together form the **chromatic scale**, which is now a complete collection of notes. The implication of this is that any further scales we derive will be a specific collection of notes taken from the chromatic scale, and that melody writing will be limited to the notes of the chromatic scale.

Note	C	C#	D	E♭	E	F	F#	G	A♭	A	B♭	B	C_{+1}
Interval Name	Unison	Minor Second	Major Second	Minor Third	Major Third	Perfect Fourth	Tritone	Perfect Fifth	Minor Sixth	Major Sixth	Minor Seventh	Major Seventh	Octave
Frequency Ratio	1	25/24	9/8	6/5	5/4	4/3	45/32	3/2	8/5	5/3	9/5	15/8	2
Equivalent Decimal	1.000	1.042…	1.125	1.200	1.250	1.333…	1.406…	1.500	1.600	1.666…	1.800	1.875	2.000

Table 2: One possible configuration of frequency ratios between notes of the (just) chromatic scale [10].

More Scales, Chords, and Their Uses

There exist further commonly used chords and scales beyond those already introduced, but their construction can be traced back to the theory we presented so far. The minor triad "i", for example, is the chord I with its pairwise frequency ratios reversed [9]. The minor scales are identical the major scale but constructed instead from i, with either iv and/or v instead of IV and V. Notice that the middle notes of these chords form three of the five extra notes in the chromatic scale above. Whilst the major triad and scales sound distinctively "happy", their minor equivalents sound "sad", because the minor triad contains additional dissonance due the interference between the fifth harmonic of the bottom note and the fourth harmonic of the middle note. Based on I, IV, and V, we can also build triads on every note of the major/minor scales, by selecting the note, its third, and fifth *note* up in that scale. This gives us seven chords to work with per major/minor scale.

Almost all Western music use either the major or minor scale for harmony writing (due to the importance of I, IV, V, and their minor equivalents) and occasionally draws additional notes from the chromatic scale for melodic decoration. How do we use these chords and scales to write music? A valid approach is to firstly write a melody using notes from a major/minor scale, then harmonise it: any given note in the melody would be accompanied by *one* of the three possible chords containing that note, or an exotic chord, such as the diminished seventh. In reality, more rules would apply regarding the melody and **chord progressions**, which are beyond the scope of this article.

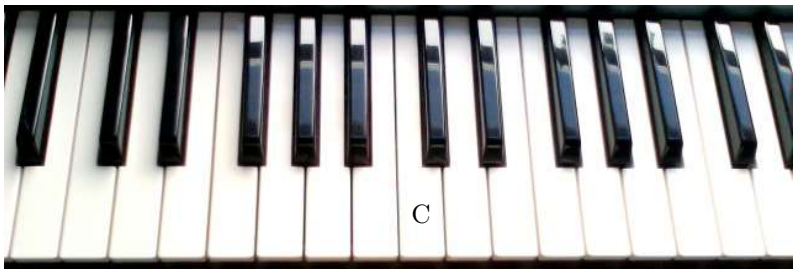

Figure 6: a piano keyboard [9], with C clearly marked. The white notes of the keyboard correspond to the major scale based on C, repeating by octaves. The black notes, in order (from C), are C#, E♭, F#, A♭, and B♭, respectively.

Tuning and Temperament

The combination of specific frequency ratios in the chromatic scale we have derived so far is named **just intonation** [10]. The name is given as we have shown that this is the best way to form a scale with consonant chords I, IV, V, and their minor equivalents, which are the most important chords in the scale. Despite its consonance, problems quickly arise when we consider its practicality for use on the piano, where we assign twelve *discrete* notes to the octave. This leads us to derive an alternative way to tune the chromatic scale.

We have so far tuned our chromatic scale based on C: all other notes have frequency ratios relative to it. If we do tune our piano in C, we can play a piece of music based, for example, on the B major scale, which is identical to only using the *chromatic* notes immediately to the left of the C major scale. This means we use the notes B, E, and the five black keys. A comparison between the C and B major scales shows that most notes in B major are quite badly out of tune, because the frequency ratios, now relative to B, are heavily distorted from the justly tuned C major scale. It can be shown [9] that several other major and minor scales are also unusable due to this distortion, some more so than others.

Note	B	C#	D# (E♭)	E	F#	G# (A♭)	A# (B♭)	B_{+1}
Frequency Ratio	1.000	1.111...	1.280	1.333...	1.500	1.706...	1.920	2.000
Just Intonation Ratio	1.000	1.125	1.250	1.333...	1.500	1.666...	1.875	2.000
Difference from Just Intonation	0%	−1.2%	+2.4%	0%	0%	+2.4%	+2.4%	0%

Table 3: comparison between the B major scale to the C major scale, in just intonation tuned on C. Note that the chord V *in B major:* F#, A# and C#, now has a different frequency relationship between the three notes, compared to I or IV.

How do we fix this issue? One approach is to own twelve pianos, each justly tuned to the notes of the chromatic scale. However, given the price of a grand piano, this is just about the least practical way to deal with the problem. Furthermore, Western music after the 18th century frequently change the major/minor scale a piece of music is based on, *within* the piece of music. The result of this, if we insist on using a piano that is justly tuned to the correct note at the start of the piece, is that the music will sound increasingly out of tune as time passes.

A far more practical solution is to abandon just intonation altogether. If we desire twelve notes, each with a constant frequency ratio that also preserves the octave, then the only solution is to tune adjacent notes using the ratio $2^{1/12}$. Since the frequency ratio between any two notes that are n notes apart in the chromatic scale is $2^{n/12}$, we see that the twelve major/minor scales are equivalent, with each being a simple transposition of the others.

This tuning system is known as **equal temperament** [10], because we *temper* with certain notes to make their frequency ratios deviate from just intonation or **Pythagorean tuning**, which is yet another tuning system which preserves the octave and eleven out of twelve just perfect fifths [10]. Equal temperament was not widely adopted until the early 19th century, where it remains the overwhelmingly universal tuning system today, in all genres of Western music. One of the first musical works which advocated towards equal temperament was J. S. Bach's (1685-1750) *The Well-Tempered Clavier*[4].

[4] In my opinion, this is the greatest composition in musical history. The dominant tuning system in Bach's time was meantone temperament, which is identical to Pythagorean temperament, but constructed from stacks of perfect fifths with a frequency ratio slightly less than an equal tempered perfect fifth, such that we obtain either eleven justly tuned major (5/4) or minor (6/5) thirds, rather than justly tuned perfect fifths as in Pythagorean temperament. Strictly speaking, well temperament is not quite equal temperament, but a very close approximation to it, otherwise the work would have been called *The Equal-Tempered Clavier*, which sounds completely terrible.

Note	C	C#	D	E♭	E	F	F#	G	A♭	A	B♭	B	C_{+1}
Just Intonation Ratio	1 = 1.000	25/24 = 1.042…	9/8 = 1.125	6/5 = 1.200	5/4 = 1.250	4/3 = 1.333…	45/32 = 1.406…	3/2 = 1.500	8/5 = 1.600	5/3 = 1.666…	9/5 = 1.800	15/8 = 1.875	2 = 2.000
Equal Temp. Ratio	1.000	1.059…	1.122…	1.189…	1.260…	1.335…	1.414…	1.498…	1.587…	1.682…	1.782…	1.888…	2.000
Difference from Just Intonation	0%	+1.7%	−0.2%	−0.9%	+0.8%	+0.1%	+0.6%	−0.1%	−0.8%	+0.9%	−1.0%	+0.7%	0%

Table 4: comparison between just intonation and equal temperament. Note that whilst the perfect fourth and fifth are very close to being in tune, the major and minor thirds are not.

Equal temperament is our best compromise on instruments with discrete notes (besides the piano, this covers virtually all woodwind instruments) because every scale is as equally usable as the others, but also just as out of tune as the others; since we see that most notes in equal temperament deviates only slightly from just intonation, we will be able to hear beats in the harmonics when we play chords. This unfortunately means that the most practical solution to our problem leads us to produce music that can never quite be "perfect". One of the more difficult and dedicate tasks for a professional orchestra is for it to carefully deviate from equal temperament at certain points within a piece of music, so the consonance provided by the other two tuning systems can be exploited.

Conclusion

The existence and common usage of the chords, scales, and temperament we have derived are beyond reasonable doubt. However, we only present one of the many ways in which these could be derived, by defining dissonance as beating and roughness. This is based on the work by Helmholtz (1821–1894), further improved by experiments in the twentieth century on the basilar membrane, which seems to be the most popular theory amongst scientists. Other theories, such as those based on treating the brain as a computational system [9], exist and they provide alternative explanations for the musical constructs. This "better theory of music" remains an active area of research, and although its followers are small in numbers, it nonetheless gives Western music a firm, scientific grounding, which, having been a musician for half my life, I find absolutely mesmerizing.

Acknowledgements

I would like to thank Liz Anya, Scott Chegg and Toby Larone for their suggestion of ideas for the article, and their generous donation of sustenance. Without their help I surely would have had to endure starvation over the New Year.

References

[1] Taylor E. *The AB Guide to Music Theory, Part I.* United Kingdom: The Associated Board of the Royal Schools of Music; 1989.

[2] Boas ML. *Mathematical Methods in the Physical Sciences* (3rd ed.). USA: John Wiley & Sons; 2005.

[3] White HE, White DH. *Physics and Music: The Science of Musical Sound.* New York: Dover Publications; 2014.

[4] Benson D. *Music: A Mathematical Offering.* Cambridge: Cambridge University Press; 2008.

[5] Shadle CH. Experiments on the acoustics of whistling. *The Physics Teacher.* 1983;21(3): 148-154. Available from: doi:10.1119/1.2341241

[6] Russell DA. *Hammer nonlinearity, dynamics and the piano sound.* Available from: https://www.acs.psu.edu/drussell/Piano/Dynamics.html [Accessed 04th January 2019].

[7] Feynman R, Leighton R, Sands M. *The Feynman Lectures on Physics Vol. I Ch. 48: Beats.* Available from: http://www.feynmanlectures.caltech.edu/I_48.html [Accessed 04th January 2019].

[8] International Organization for Standardization. ISO 16:1975. *Acoustics -- Standard tuning frequency (Standard musical pitch).* Switzerland: ISO; 1975.

[9] Wilkerson D. *Harmony Explained: Progress Towards A Scientific Theory of Music.* 2014. Available from: arxiv.org/html/1202.4212 [Accessed 04th January 2019].

[10] Barbour J. *Tuning and Temperament: A Historical Survey.* USA: Michigan State College Press; 1951.

[11] Wolfe J, Smith J, Tann J, Fletcher NH. *Modern B Flute Acoustics.* Available from: http://newt.phys.unsw.edu.au/music/flute/modernB/A4.html [Accessed 04th January 2019].

[12] Wolfe J. *Clarinet Acoustics - B4.* Available from: http://newt.phys.unsw.edu.au/music/clarinet/B4.html [Accessed 04th January 2019].