Emergence of collective intonation in the musical performance of crowds

Lacasa, L

Emergence of collective intelligence in musical performances

Lucas Lacasa*
School of Mathematical Sciences, Queen Mary University of London, Mile End Road E14NS London (UK)

The average individual is typically a mediocre singer, with a rather restricted capacity to sing a melody in tune. Yet when many singers are assembled to perform collectively, the resulting melody of the crowd is suddenly perceived by an external listener as perfectly tuned -as if it was actually a choral performance- even if each individual singer is out of tune. This is an example of the so-called wisdom of crowds effect and can be routinely observed in music concerts or other social events, when a group of people spontaneously sings at unison. In this paper we provide a simple mechanistic explanation for the onset of this collective phenomenon.

The wisdom of crowds [1] is a popular concept englobing several examples of collective intelligence, that emerges where the collective response of a group of entities is in some sense better than individual ones. Pioneered by Galton [2], this effect was in its simpler incarnation a direct consequence of the law of large numbers. Evidence of collective intelligence spans today social systems in different species [3–5] and activities ranging from optimal estimation [3], navigation [6], or sensing [7] to cite a few. In this work we focus on the phenomenon of collective musical performance. We are not interested in choral performances but on self-organized 'crowd performances' that take place in popular music concerts [8], sport events (e.g. in football stadiums) or other social events, or simply within groups of people that join together to perform a song or melody. Our contention is that whereas the average individual is not necessarily a gifted performer and does not particularly sing *in tune* (i.e. individual musical performances are typically of poor quality), when a large group of these imperfect singers perform at unison the resulting collective signal is surprisingly *tuned*. As a consequence, crowd performance is enhanced as compared to individual ones and is thus perceived as a choral one. Whereas some research suggest that individuals improve while performing at unison [9] -suggesting that imitation might be underpinning this phenomenon-, here we show that imitation, while clearly boosting this effect, is not itself required for the enhancement to take place in the first place. We present a toy model that supports this claim and that provides a simple explanation for the origin of this collective phenomenon.

In our toy model the melody to be played consists of a pure tone with frequency T, which the crowd interprets at unison. As individuals (or agents, from now on) don't usually have a perfect ear, the collective output produced by the crowd will be a inharmonic complex tone that develops out of the mixture of each agent's contribution. For simplicity we will assume that each agent contributes with a pure tone -i.e. a sinusoidal wave with a single frequency-. Consider two tones with frequency f_1 and f_2 and similar amplitudes, played simultaneously. What is the pitch of this complex tone? If f_1 and f_2 are sufficiently close, then the pitch is somewhere close to $(f_1 + f_2)/2$ and is accompanied with a beating at $|f_1 - f_2|$. As their difference increases the beating disappears, and for sufficiently different frequencies one can indeed perceive both frequencies. Furthermore, if $f_2 = pf_1$ for some integer p, then the pitch of the complex tone is just f_1, coinciding with the fundamental frequency that corresponds to the greatest common divisor between both frequencies, GCD(f_1, f_2). For a complex tone with $N > 2$ partials, the story is even more intricate. The fundamental frequency can be easily worked out as GCD$(f_1, f_2, \ldots, f_N)$. If we superpose the frequencies $f, 2f, 3f, ...$ (that is, a fundamental f and a few higher harmonics) with more or less the same amplitude, then the resulting pitch will indeed be f, coinciding again with the fundamental frequency. If we now superimpose $2f, 3f, 4f, ...$ we will still perceive f which in this case is indeed a missing partial. This happens as in the range 20-2000Hz, the ear has the ability to fuse harmonically-related frequencies into a single entity with a fundamental frequency, even in the case where such fundamental frequency is missing. Now, in general the fundamental frequency (either physically present or missing) does not correspond to the *perceived pitch*, i.e., to the effective frequency perceived by an external agent who is listening to the collective output. As a matter of fact, pitch is a psychoacoustic phenomenon more complex than basic frequency superposition. This fact becomes evident when we combine partials which are not harmonically related. Consider the mixture of frequencies at 120, 220, 320, 420, 520, and 620 Hz in equal measure. The GCD of the mixture is 20 Hz (a frequency which is indeed barely audible), however the perceived -or virtual- pitch coincides in this case with a mysterious frequency of 104.6 Hz [10].

There are essentially two theories (or groups of theories) on pitch perception [10], namely those focusing on spatial separation of partials in the ear (Fourier decomposition theories pioneered by Ohm and Helmholtz) and those that focus on the temporal separation, pioneered by Licklider. These are not necessarily mutually exclu-

*Electronic address: l.lacasa@qmul.ac.uk

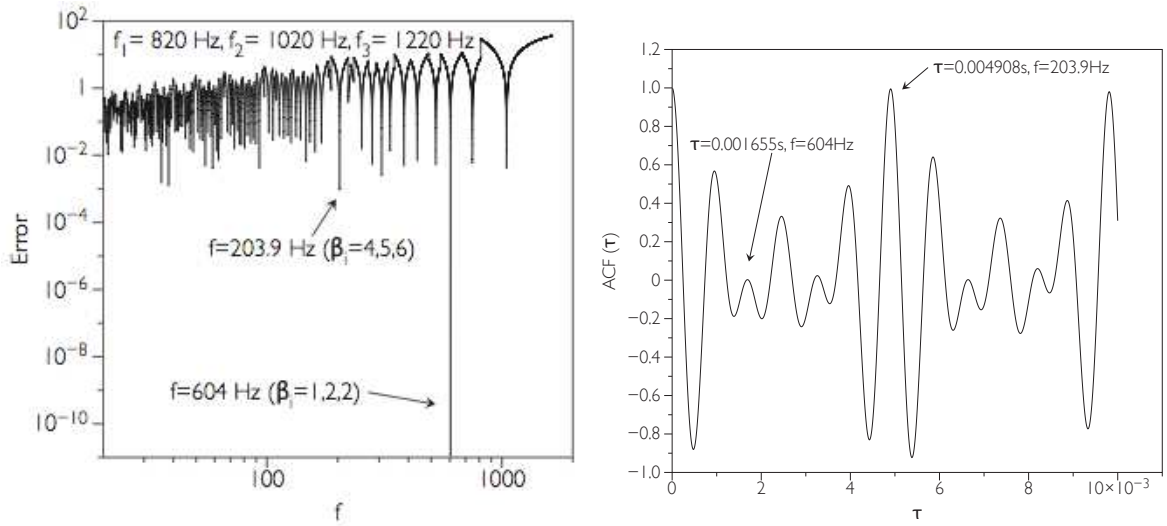

FIG. 1: (Left panel) Numerical evaluation of the roots of equation 1 for a complex tone of three partials with $f_1 = 820$ Hz, $f_2 = 1020$ Hz, $f_3 = 1220$ Hz. Each solution -denoted as a frequency with low numerical error- corresponds to a local peak in the autocorrelation function (and its harmonics). Two possible solutions with different β_i are depicted. The perceived pitch is indeed $\bar{f} = 203.9Hz$ [10, 14], which corresponds to $\beta_i = 4, 5, 6$. (Right panel) Autocorrelation function of the same complex tone, where one can appreciate that the perceived pitch is indeed associated with the first non-trivial 'large peak', whereas other peaks that take place sooner are not strong enough to develop into the perceived pitch.

sive, and both might have their range of validity. Here we use the latter approach as this seems to be better suited to deal with the range of frequencies usually displayed in modern music. As pitch in this context is related to the tendency to find repetitions at given intervals, it makes sense to look at the presence of these repetitions over time in the autocorrelation function. For instance, in the example above the first non-trivial peak occurs at $t = 0.009565s$ which is indeed related to a periodic repetition at $f = 104.6$ Hz, precisely equal to the perceived pitch.

To be more precise, let us consider a complex tone of N sinusoidal partials with frequency f_i and amplitude a_i given by $s(t) = \sum_{i=1}^{N} a_i \sin(2\pi f_i t)$, and denote by $\bar{f}$ the perceived pitch of this mixture. Then $\bar{f} = 1/\tau_M$ where τ_M is the time position of the earliest tall peak in the autocorrelation function $C(\tau) = \langle s(t)s(t+\tau)\rangle_t$. This is an extremum thus $dC/d\tau|_{\tau_M} = 0$. Moroever, as the product $\sin(t)\sin(t+\tau)$ is maximized for τ being a multiple of 2π, it is easy to see that for any local peak at τ_M of the autocorrelation function one has $\sin(2\pi f_i \tau_M) \approx 2\pi(f_i\tau_M - \beta_i)$, for some integer β_i. Putting all these conditions together, according to Heller [10] the peaks of the autocorrelation function fulfil the following self-consistent equation

$$1/\tau_M = \bar{f} \approx \frac{\sum_{i=1}^{N} a_i^2 f_i^2}{\sum_{i=1}^{N} a_i^2 \beta_i f_i}, \qquad (1)$$

where for $i = 1, \ldots, N$, $\beta_i \in \mathbb{Z}$ is the nearest integer to $f_i/\bar{f}$. This formula was first derived in the context of molecular spectroscopy to account for the so-called missing mode effect (MIME) [11–13]. It is important to highlight that $\bar{f}$ is not just a convoluted average of each frequency [12], but in some sense is an emergent quantity out of the combination of partials, much like in the luminescence spectra of complex molecules some regularly spaced vibronic progressions emerge even if they don't correspond to any ground-state normal mode of vibration (or average) of the molecule [13]. We solve eq.1 numerically and assume that $f_{\rm app}$ is a good approximation to $\bar{f}$ if

$$\left(f_{\rm app} - \frac{\sum_{i=1}^{N} a_i^2 f_i^2}{\sum_{i=1}^{N} a_i^2 \hat{\beta}_i f_i} \right) \cdot 100/f_{\rm app} < \epsilon,$$

where $\hat{\beta}_i$ is the nearest integer to $f_i/f_{\rm app}$. As a rule of thumb, we set $\epsilon = 10^{-2}$, which means that $f_{\rm app}$ satisfies eq. 1 self-consistently with an error which is less 0.01% of the frequency $f_{\rm app}$.

Note that eq.1 is multivalued: at least it admits as solutions the virtual pitch and its infinitely many subharmonics. Indeed, for equal frequencies $f_i = k\ \forall i$, the virtual pitch is trivially $\bar{f} = k$ and this is a solution of eq.1 for $\beta_i = 1\ \forall i$, but so are subharmonics k/p, for $p \in \mathbb{N}^+$ and $\beta_i = p\ \forall i$. Also, for two close enough frequencies $f_1 \approx f_2$, then $(f_1 + f_2)/2$ is an approximate solution which indeed corresponds to the perceived pitch. Now, eq.1 captures the location of peaks in the autocorrelation function, but unfortunately not their height. Consider for instance the complex tone formed by partials of equal amplitude at frequencies 820, 1020, and 1220 Hz. The GCD is 20 Hz, right at the threshold of hearing, and seems an unlikely perceptual result of combining these much higher frequencies. Pierce [14]

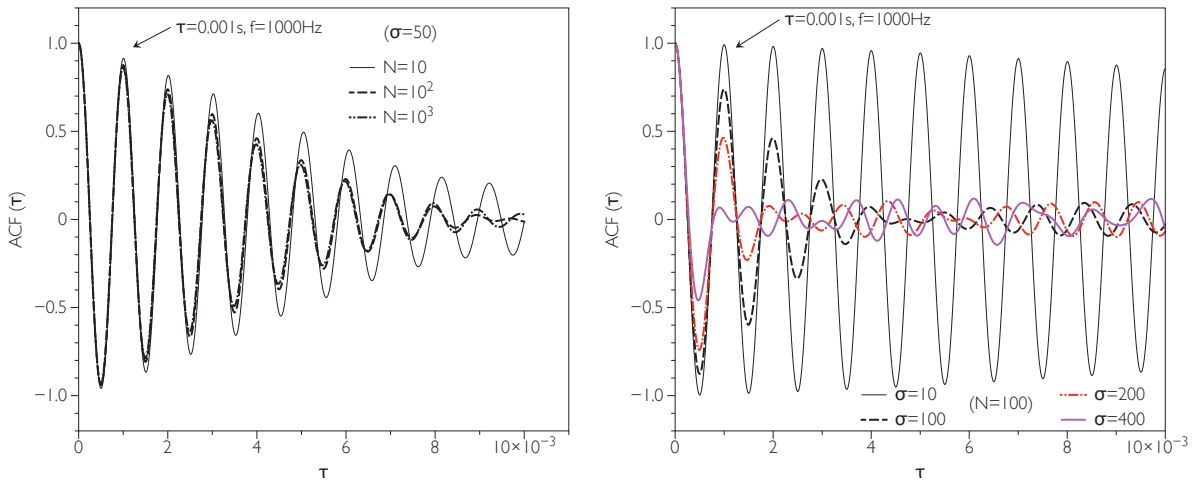

FIG. 2: (Left panel) Autocorrelation function of a complex tone formed by N frequencies $f_i \sim \mathcal{N}(1000, 50)$, for $N = 10$ (solid line), $N = 10^2$ (dashed line) and $N = 10^3$ (dashed dotted line). In every case the first non-trivial large peak in the autocorrelation function lies at $tau = 0.001$ seconds, yielding a perceived pitch $\bar{f} = 1000Hz$. Interestingly, other peaks in the autocorrelation function (associated with other solutions of eq. 1) vanish as N increases with a where one can appreciate that the perceived pitch is indeed associated with the first non-trivial 'large peak', whereas other peaks that take place sooner are not strong enough to develop into the perceived pitch. (Eight panel) Autocorrelation function of a complex tone formed by $N = 100$ frequencies $f_i \sim \mathcal{N}(1000, \sigma^2)$, for increasing values of σ^2. The perceived pitch converges to $\bar{f} = 1000Hz$ for a rather large range of values σ^2, after which the complex tone does not have a clear perceived pitch.

cites this case as an interesting example and reports that the perceived pitch is 204 Hz. A possible solution can be found for $\beta_1 = 1, \beta_2 = \beta_3 = 2$, for which setting $a_i = 1$, we get $\bar{f} = 604$ Hz. According to the left panel of fig.1, this seems indeed the solution with minimal numerical error. The solution with second minimal error corresponds to a higher combination $\beta_1 = 4$, $\beta_2 = 5$, $\beta_3 = 6$ for which $\bar{f} \approx 203.9$ Hz. However, it is this latter candidate that coincides with the empirical value found by Pierce. If we look at the autocorrelation function of the complex tone (right panel of the same figure), we indeed discover peaks at $1/604 = 0.001655$ and $1/203.9 = 0.004908$ seconds (among others), however the latter is the sharpest peak and hence constitutes the perceived pitch.

All in all, the systematic computation of the perceived pitch is not straightforward. Heller [10] speaks about three criteria to determine what peak corresponds to the perceived pitch of a complex tone: (i) the sooner in the autocorrelation function (sooner times corresponds to larger frequencies), (ii) the larger the autocorrelation of the peak, and (iii) the sharper the peak. However looking at the solutions of eq.1 we are only able to discern criterion (i), therefore in what follows we will focus on the autocorrelation function to discern the perceived pitch from the set of solutions of eq.1.

Basic model. Consider N agents aiming to sing at unison a given frequency T. We assume that all agents sing pure tones (i.e. sinusoids of frequency f_i) at approximately the same amplitude ($a_i = K\ \forall i$ for some $K \in \mathbb{R}^+$) and model the imperfection of each agent as an independent Gaussian deviation. That is, $\forall i = 1, \ldots, N$ the frequency $f_i = f^t + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2)$. The standard deviation σ therefore tunes the diversity of imperfections. Note that trivially, $\lim_{N\to\infty} \mathrm{GCD}(f_1 \ldots f_N) = 0$. Is there a perceived pitch for this complex tone? Applying eq.1 in this case, one finds a frequency

$$\bar{f} \approx \frac{\sum_{i=1}^N (T + \xi_i)^2}{\sum_{i=1}^N \beta_i (T + \xi_i)} = \frac{\sum_{i=1}^N T^2 + \sum_{i=1}^N \xi_i^2 + 2\sum_{i=1}^N T\xi_i}{\sum_{i=1}^N \beta_i T + \sum_{i=1}^N \beta_i \xi_i}.$$

To prove that the crowd sings better than each individual in a nontrivial way, we need to (i) find that $\bar{f} \approx T$ is a solution to the latter equation that (ii) corresponds to an early tall peak in the autocorrelation function of the signal and that (iii) this holds for a range of values of σ. The criterion (iii) simply means that if the phenomenon only holds for very small σ, one could argue that in practice the perceived pitch harmonically fuse barely audible deviations from the correct pitch. In the contrary, if σ is large enough such that every random sample is almost surely out of tune then the emergence of a tuned perceived pitch will be clearly a collective effect.

First, as eq.1 is multivalued we focus in the solution associated to $\beta_i = 1\ \forall i$. In this case, trivially $\sum_{i=1}^N T^2 = NT^2$ and $\sum_{i=1}^N \beta_i T = NT$. According to the central limit theorem, the sum of N Gaussian random variables $\mathcal{N}(0, \sigma^2)$ variables is a Gaussian random variable $\mathcal{N}(0, N\sigma^2)$. Thus for $N \gg 1$ we can use expected values such that $\sum_{i=1}^N \beta_i \xi_i \to 0$ and $\sum_{i=1}^N \xi_i^2 \to N\langle \xi^2 \rangle = N\sigma^2$ (alternatively, the sum of N squared standard Gaussian random variables is a random variable which is distributed as a χ^2 distribution with mean N, so if the orig-

inal Gaussian variables are not standard but have variance σ^2, then the mean of the rescaled χ^2 distribution is $N\sigma^2$). Altogether, the solution to eq.1 associated to $\beta_i = 1\ \forall i$ is

$$\bar{f} \approx \frac{NT^2 + N\sigma^2}{NT} = T + \sigma^2/T \qquad (2)$$

Provided that extremely large deviations from the correct tone are not abundant among individual performance (so that $\sigma \ll T$ is a good approximation) then the second term in the latter solution is $\ll T$ and then at leading order $\bar{f} \approx T$. Now, to evaluate whether this frequency indeed corresponds to the earliest tall peak in the autocorrelation funcion, we have tested this prediction numerically in figure 2. While human hearing ranges from 20 to 20000 Hz, the greater sensitivity is known to lie within 200 and 2000 Hz. We therefore discard solution frequencies under 100 Hz (that is, times larger than 10^{-2} seconds) as they will only contribute to perceived background noise. In the left panel of figure 2 we plot the autocorrelation function of a complex tone made by N sinusoids with equal amplitude and frequencies $f_i \sim \mathcal{N}(T, 50)$, for $T = 1000$ Hz. We can observe that as the number of agents N increases, the frequency $\bar{f} = T$ indeed emerges as the clear perceived pitch (numerical evaluation of the solutions of eq.1 are plotted in an appendix figure). In the right panel of the same figure we explore the effect of increasing σ in the shape of the autocorrelation function for $N = 100$ frequencies (solutions of eq.1 in this case are again summarized in an appendix). As expected, the frequency $\bar{f} = 1000$ Hz coincides with the perceived pitch for a reasonably large range of values of σ. For $\sigma = 10$ and 100 the peak is clearly visible although it decreases as σ increases. Note that the just noticeable difference (which quantifies the threshold at which a change in pitch is perceived) depends on the frequency and for 1000 Hz this is smaller than 10 Hz. This means that for $\sigma = 100$ most of the individual agents will be effectively out of tune, however the perceived pitch of the aggregate will still emerge as being in tune. Accordingly, the emergent pitch is robust even if each agent is not particularly gifted, musically speaking. For even larger values ($\sigma = 400$) the spectrum approaches a flat shape, the peak has fade away and no clear pitch emerges accordingly. All in all, we can conclude that a crowd indeed sings better as a whole than each individual separately, even if no synchronization takes place among individuals.

Introducing short-range interactions. Remarkably, our basic model does not require that each agent interacts for the perceived pitch to emerge as the 'correct' collective intonation. It is however true that in realistic cases individuals that sing in groups tend to tune up with their surrounding, if in their close neighborhood there is at least some other person with better intonation [9]. The intonation and imitation capacities are definitely heterogeneous across people,

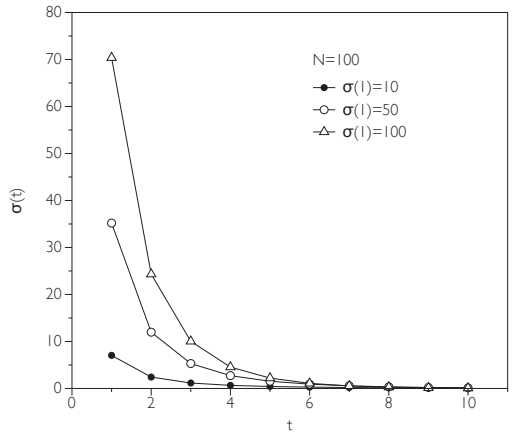

FIG. 3: Numerical evaluation of standard deviation of the distribution of frequencies of the lattice model, as interactions take place over time. Already after one simulation step, the effective standard deviation considerably decreases, what in turn implies that the virtual pitch approximates its leading order $\bar{f} \approx T$

in what follows we show that in general terms this imitation process effectively reduces the value of σ in eq.2.

To explore the effect of imitation we propose the following toy model: we locate agents in the vertices of a two-dimensional lattice, and make them vibrate at a given frequency $f_i^{(t)}$ which can now be dynamically updated. Initially we again set $f_i^{(0)} = T + \xi$, for Gaussian i.i.d. random variables $\xi \sim \mathcal{N}(0, \sigma^2)$. Then, at each simulation time t the dynamics for each agent i are such that: (i) if any surrounding agent is singing more *in tune* than i (modeled by the fact that in the Von Neumann neighborhood of i $|f_i^{(t)} - T| < \min\{|f_j^{(t)} - T|\}_{\rm nn}$), then (ii) i updates his frequency by imperfect imitation, such that $f_i^{(t+1)} = f_i^{(t)} + C(i)[\min - f_i^{(t)}]$, where min is the frequency to be imitated and $C(i) \in [0, 1]$ is a real number that describes the fitness of agent i to imitate or tune up. Intuitively, an agent with good imitation skills will initially perform close to T, so for simplicity we define $C(i) = 1 - \frac{|f_i^{(0)} - T|}{T}$.

This process is then run in parallel and iterated in time. The relevant observable of the system is again the perceived pitch $\bar{f}(t)$ which is now a function of time and will change as the frequencies variance $\sigma^2(t)$ is modified. If imitation is null $C(i) = 0$, then this model reduces to the non-interacting case above. At the other extreme, if every agent has perfect imitation skills $C(i) = 1$, then there is an absorbing state where all the agents end up vibrating at the same characteristic frequency f_i^* that corresponds to the one for which $|f_i^{(0)} - T|$ is minimized. That is to say, amongst the initial values of the partials, the one closest to T percolates and emerges as a consensus. In this ideal situation, it is easy to see that as $N \to \infty$, $\bar{f}(\infty) \to T$. In the more realistic case

of bounded capacity $C(i) \in (0,1)$, the absorbing state will be such that $\xi_i^{(\infty)}$ will not just have one value but several (corresponding to several degrees of intonation). However what is straightforwardly guaranteed is that the variance of the frequencies distribution σ^2 will decrease over time with respect to the initial condition (non-interacting case). That is to say, the second term is eq.2 will be smaller as interactions take place, boosting even further the wisdom of crowds effect. These tendencies are confirmed by numerical simulations in figure 3.
As a final comment, note that in the event that the crowd is the audience of a concert which follows the band's lead singer (i.e. the system is coupled to an external 'pitch field'), then the imitation process directly takes place with the singer, instead of locally. This mechanism trivially uncouples the system and reduces the problem to the original non-interacting case, albeit with a new σ which is much smaller than in the original case.

To conclude, we have given a simple explanation for the onset of collective intelligence in crowds that sing at unison. Within reasonable limits, regardless the intonation of each singer the collective tone will be perceived as to be in tune. This wisdom of crowds effect is further boosted if one allows individuals to adjust their frequency by any degree of imperfect imitation with his neighbors, although this additional mechanism is not required for the collective effect to emerge in the first place. Interestingly, this result does not require subjects to follow any leader, and emerges in a self-organized way due to the psychoacoustic properties of the perceived pitch.

Acknowledgments

I thank Andrew Berdahl for fruitful discussions and encouragement.

[1] J. Surowiecki, The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations (Doubleday, 2004).
[2] F. Galton, Vox Populi. *Nature* **75**, 450 (1907).
[3] E. Bonabeau, M. Dorigo and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems* (Oxford University Press, NY, 1999)
[4] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, T. W. Malone, Evidence for a collective intelligence factor in the performance of human groups. *Science* **330**, 686 (2010).
[5] N. R. Franks, S. C. Pratt, E. B. Mallon, N. F. Britton, D. J. Sumpter, Information flow, opinion polling and collective intelligence in house-hunting social insects. *Phil. Trans. R. Soc. B* **357**, 1567 (2002)
[6] A.M. Simons, Many wrongs: the advantage of group navigation, *Trends in Ecology & Evolution* **19**, 9 (2004)
[7] A. Berdahl, C.J. Torney, C.C. Ioannou, J.J. Faria, and I.D. Couzin, Emergent sensing of complex environments by mobile animal groups, *Science* **339**, 6119 (2013).
[8] Example that shows that within a crowd, individuals perform poor https://m.youtube.com/watch?v=Y0bpUCVt6hE but the crowd performs beautifully https://m.youtube.com/watch?v=AWaItQhFnRQ.
[9] G.A. Green, Unison versus individual singing and elementary students vocal pitch accuracy, *Journal of Research in Music Education* 42, 2 (1994).
[10] E.J. Heller, *Why You Hear What You Hear: An Experiential Approach to Sound, Music, and Psychoacoustics* (Princeton University Press 2012) ISBN: 9781400845583
[11] L. Tutt, D. Tannor, E. J. Heller and J.I. Zink, The MIME effect: absence of normal modes corresponding to vibronic spacings, *Inorg. Chem.* **21** (1982) 3858-3859
[12] L. Tutt, D. Tannor, J. Schindler, E. J. Heller and J.I. Zink, Calculation of the missing mode effect frequencies from Raman intensities, *J. Phys. Chem.* **87** (1983) 3017-3019
[13] L.W. Tutt, J.I. Zink and E.J. Heller, Simplifying the MIME: a formula relating normal mode distortions and frequencies to the MIME frequency, *Inorg. Chem.* **26** (1987) pp.2158-2160
[14] John R. Pierce *The Science of Musical Sound* (Scientific Academic Library, 1992).

Appendix

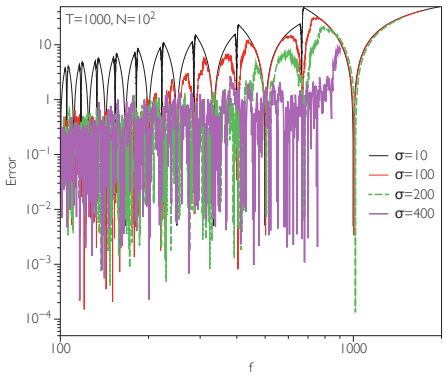

FIG. 4: Numerical evaluation eq.1 for $N = 10^2$ frequencies $f_i \sim \mathcal{N}(T = 1000, \sigma^2)$. For $\sigma \ll T$, there are few solutions that consist of T and its subharmonics. As σ increases, other solutions start to appear, and $\bar{f} = T$ eventually disappears.

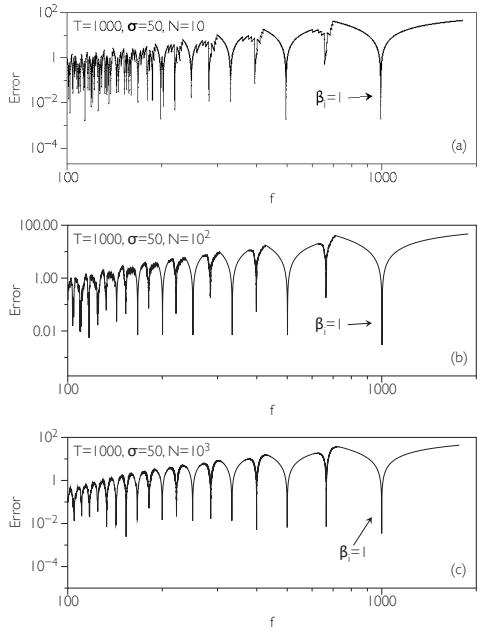

FIG. 5: Numerical evaluation of the roots of equation 1, where frequencies $f_i \sim \mathcal{N}(1000, 50)$ for $N = 10, 10^2$ and 10^3 respectively. As N increases, just a few frequencies (and subharmonics) emerge as the numerical solutions.