The Pitch of Vibrato Tones

A Model Based on Instantaneous Frequency Decomposition

Bruno A. Mesz and Manuel C. Eguia

Laboratorio de Acústica y Percepción Sonora, Universidad Nacional de Quilmes, Buenos Aires, Argentina

We study vibrato as the more ubiquitous manifestation of a nonstationary tone that can evoke a single overall pitch. Some recent results using nonsymmetrical vibrato tones suggest that the perceived pitch could be governed by some stability-sensitive mechanism. For nonstationary sounds the adequate tools are time–frequency representations (TFRs). We show that a recently proposed TFR could be the simplest framework to explain this hypothetical stability-sensitive mechanism. We propose a one-parameter model within this framework that is able to predict previously reported results and we present new results obtained from psychophysical experiments performed in our laboratory.

Key words: **pitch perception; time–frequency analysis; vibrato**

Introduction

The vast majority of studies on pitch perception have been focused on stationary stimuli, primarily complex tones, formed by many steady frequencies. However, in natural sounds and also in music, the frequency components usually vary in time in a coherent way. Harmonic complex tones, for example, have a well-defined fundamental frequency (F0) that is strongly correlated with pitch for steady tones. When these tones are produced in a natural context, the F0 is never constant and the pitch assignment is less clear. Still, under certain circumstances, we are able to perceive a single pitch for a tone whose F0 is fluctuating in time. A foremost example of this is vibrato in music, for which some recent results using nonsymmetrical vibrato tones suggest that the perceived pitch could be governed by some stability-sensitive mechanism.[1] For nonstationary sounds the adequate tools are time–frequency representations (TFRs). We show that a recently proposed TFR[2] could be the simplest framework to explain this hypothetical stability-sensitive mechanism.

Vibrato is a widespread performing technique produced by a quasi-periodic modulation of the frequency components of a note that gives a single overall pitch, despite the fact that the modulation range can surpass the semitone. In contrast to the common belief that vibrato can be used to mask poor intonation, carefully designed psychophysical experiments have shown that the pitch can be extracted with a limen that is less than one-tenth of semitone.[3] However, the pitch value that is assigned to vibrato tones has been a matter of debate.

Shonle and Horan[4] reported that the geometric mean of F0 was a good approximation to the assigned principal pitch. A similar conclusion was drawn by Iwamiya *et al.*[5] However, d'Alessandro and Castellengo[6] found that the later parts of the modulation were more important for the pitch judgment. Shonle and Horan[4] also reported that for nonsymmetrical modulations the principal pitch was shifted from the geometric mean of peaks and troughs and

Address for correspondence: Manuel C. Eguia, Laboratorio de Acústica y Percepción Sonora, Universidad Nacional de Quilmes, R. S. Peña 352 Bernal, 1876 Buenos Aires, Argentina. meguia@unq.edu.ar

The Neurosciences and Music III—Disorders and Plasticity: Ann. N.Y. Acad. Sci. 1169: 126–130 (2009).
doi: 10.1111/j.1749-6632.2009.04767.x

suggested a nontrivial time-averaging of the frequencies present in the stimulus. Gockel *et al.*[1] confirmed these pitch shifts for nonsymmetrical modulated sinusoids and proposed a model where the frequency averaging is weighted by a function that decreases when the rate of modulation increases. In other words, if the modulation has portions of slow and fast variation of F0, the slow parts contribute more to the overall principal pitch than the fast ones. The authors termed this mechanism "stability-sensitive weighting" and suggested that the "pitch estimate for a given segment of the stimulus may be reduced in accuracy when the frequency changes rapidly during a segment" (p. 702).[1]

In this work, we will show that this proposed stability-sensitive mechanism can be derived from a nonlinear TFR recently proposed by Gardner and Magnasco.[2,7] This method, in turn, belongs to the reassignment class of TFRs and provides an accurate representation of rapidly changing sounds (and vibrato in particular). In fact, the core of the method is the notion of "consensus," which computes the consistency of the reassignments of time and frequency. Our working hypothesis is that the stability-sensitive mechanism described in Gockel *et al.*[1] is a natural outcome of the consensus criterion defined in Ref. 2.

In order to test our hypothesis we developed a simple model based on the reassignment method and a consensus measure, and performed a psychophysical experiment that explored parametrically the transition between a symmetrical modulation and a highly nonsymmetrical one.

Experiment

Our stimuli consisted of frequency-modulated sinusoids. We used trapezoidal modulation profiles, each cycle consisting of a flat portion (constant frequency) and two linear segments forming a peak. The percentage of the cycle occupied by the flat portion (**p**) was used as a parameter. The peaks could be pointing toward the high frequencies (which we called the **u** profile) or toward the low frequencies (**n** profile). For $\boldsymbol{p} = 0$ the modulation shape is triangular and there is no distinction between **u** and **n** profiles.

For all stimuli we used four cycles of modulation with a geometric mean of 1000 Hz, a modulation rate of 10 Hz, and an overall depth of 150 cents. The starting phase of the modulation was varied randomly at each presentation in order to wash out possible start or end effects.

Four subjects participated in the experiment. All were normal-hearing individuals with a varying degree of musical training and ages ranging from 25 to 38 years. We divided the experiment session in seven blocks, one for each type of modulation: $p = 0$ (symmetrical) and $p = 25$, 50, 75 for the **u** profiles and **n** profiles. Each block consisted in obtaining four pitch matches between the modulated tone and an unmodulated sinusoid. A two-interval, two-alternative forced choice and an adaptive staircase procedure were used to obtain the pitch matches.

Results

Figure 1 displays the mean results of our experiment. The values presented are the mean adjusted frequencies (averaged over the subjects) along with their standard errors, for the seven stimuli presented. In the case of the symmetrical modulation ($p = 0$) the mean adjusted frequency is close to the geometric mean, as expected. For nonsymmetrical modulations we obtained pitch shifts consistent with previous results: the mean adjusted frequencies were closer to the flat portion of the modulation.

Model

We now move on to the formulation of a model of principal pitch based on the

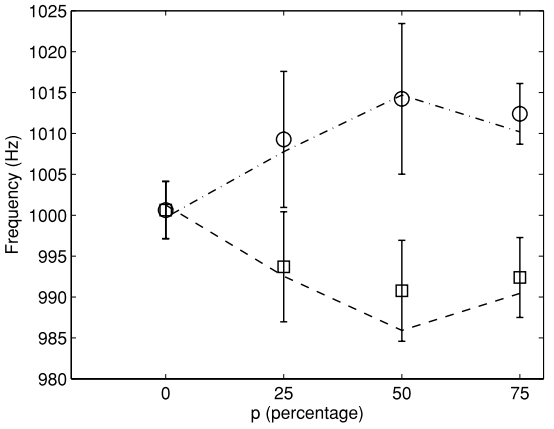

Figure 1. Pitch shifts observed for the perceived principal pitch of the modulated tones described in the text. For the **u** profiles (squares) the four subjects reported principal pitches that were lower than the geometric mean of 1 kHz. The tones with **n** profiles (circles) evoked pitches that were higher than the mean. The prediction of our model (see Eq. 7) for a parameter value of $\gamma = 0.04$ and for **u** profiles (**n** profiles) is plotted in dot-dashed lines.

reassignment method. This model relies mainly on phase information obtained from the acoustic signal which is well preserved in the peripheral auditory system. The model has three stages: (1) instantaneous frequency extraction, (2) cross-check across channels, and (3) amplitude weighting.

We start with the short-time Fourier transform (STFT) of the acoustic signal $x(t)$ computed at a discrete set of frequencies ω, referred to as channels. The STFT can be rewritten in terms of its amplitude X and phase Φ:

$$\begin{aligned} STFT(t,\omega) &= X(t,\omega)\, e^{\Phi(t,\omega)} \\ &= \int x(t+\tau)\, h(-\tau)\, e^{-i\omega t}\, d\tau \end{aligned} \tag{1}$$

We define the *channelized instantaneous frequency* (CIF) as the time derivative of the STFT phase:

$$CIF(t,\omega) = \frac{\partial}{\partial t}\Phi(t,\omega) \tag{2}$$

Then we estimate the instantaneous frequency of the signal as the average of the CIFs present across all channels, weighted by the energy of the channel (computed as the square of X).

$$FI(t) = \frac{\sum_{\omega} CIF(t,\omega)\, X(t,\omega)^2}{\sum_{\omega} X(t,\omega)^2} \tag{3}$$

In the second stage, we compare the CIFs computed at neighboring channels. The idea is to assign to the stimulus a time-varying perceptual weight proportional to the degree of similarity among the instantaneous frequency estimates in different channels. This is equivalent to the consensus introduced in Ref. 2.

The simplest way to compare the CIFs across channels is to compute the derivative of this quantity respect to the channels (ω). Since the CIF is the time derivative of the phase Φ, this magnitude is equivalent to the mixed partial derivative of the phase, and we term it MPD in the following:

$$\begin{aligned} MPD(t,\omega) &= \frac{\partial}{\partial\omega}\left(\frac{\partial}{\partial t}\Phi(t,\omega)\right) \\ &= \frac{\partial}{\partial\omega} CIF(t,\omega) \end{aligned} \tag{4}$$

By the very definition of a derivative, the modulus of the MPD measures the rate of change of the CIFs across channels. A small (respectively, big) MPD modulus can then be regarded as indicative of high (respectively, low) consensus. These considerations led us to consider a perceptual weighting function W1 that has a maximum value associated to maximum consensus and decays with decreasing consensus. This weighting function is also averaged proportionally to the energy of each channel:

$$\begin{aligned} &W1(t;\gamma) \\ &= \frac{\sum_{\omega} X(t,\omega)^2 \exp\left(-|MPD(t,\omega)|/\gamma\right)}{\sum_{\omega} X(t,\omega)^2} \end{aligned} \tag{5}$$

We afterward use γ as our only fitting parameter. Note that for γ small, the weight W1 tends to emphasize the points where the MPD is close to zero and to neglect the others, while for

TABLE 1. Previously Reported Results of the Perceived Principal Pitch of Nonsymmetrical Vibratos[1,4]

Modulation shape	CF (Hz)	FM depth (cents)	AM (degree)	FM rate (Hz)	Starting phase (radians)	Mean (Hz)	Std (Hz)	Prediction (Hz)	Reference
Trapezoidal ∪	368	100	0	6	0	362	14	362	Shonle and Horan[4]
Trapezoidal ∩	524					546	16	538	
Nonsymmetrical ∪	1000	135	0	10	π	989.6	2.6	989.8	Gockel *et al.*[1]
					0	990.0	2.2	988.7	
Nonsymmetrical ∩					π	1009.7	2.1	1010.7	
					0	1011.2	2.0	1011.7	

The center frequency (CF) is the geometric mean of the modulation. The prediction of our model was made with the same parameter value ($\gamma = 0.04$) that fits our experimental results. We also adjusted the results reported in Ref. 5 (data not included).

higher values of γ the weight gives equal relevance to all MPD values.

In order to include vibratos that also have amplitude modulation, we add a second weight W2 equal to square root of the energy sum across channels:

$$W2(t) = \sqrt{\sum_{\omega} X(t, \omega)^2} \tag{6}$$

Finally, the principal pitch (PP) is computed averaging over discrete times:

$$PP(\gamma) = \frac{\sum_{t} FI(t)\, W1(t;\gamma)\, W2(t)}{\sum_{t} W1(t;\gamma)\, W2(t)} \tag{7}$$

We examined the applicability of the model to previous results on principal pitch perception of vibrato tones.[1,4,5] Despite the diversity of psychophysical experiments and the inclusion of amplitude modulation in Ref. 5, we were able to adjust all the results within the reported standard deviation using a single parameter value $\gamma = 0.04$ (Table 1). Actually there is a wide range of γ values that can fit the data reasonably well. The prediction with the same parameter value ($\gamma = 0.04$) for our experiments is displayed in Figure 1 as dashed lines and we again observe good agreement.

Conclusions

We presented a new model for principal pitch perception of vibrato tones that can predict previously published results. The proposed model has some biological plausibility since the local computations at each auditory channel needed for pitch estimation are based on phase information, so they could be implemented in principle using the time interval between action potentials in the auditory nerve fibers. There is a long-standing debate regarding whether pitch perception is more related to place-rate coding on the tonotopic axis or to temporal coding. Our model integrates tonotopic organization (related to the channel amplitudes) and temporal coding (through the CIFs) and a cross-check among channels of both kind of information (consensus) in order to give a single pitch percept.

Acknowledgments

We thank M. Magnasco for useful discussion. This work was partially supported by the FONCyT.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Gockel, H., B.C.J. Moore & R.P. Carlyon. 2001. Influence of rate of change of frequency on the overall pitch of frequency-modulated tones. *J. Acoust. Soc. Am.* **109:** 701–712.
2. Gardner, T.J. & M.O. Magnasco. 2006. Sparse time-frequency representations. *Proc. Natl. Acad. Sci. USA* **103:** 6094–6099.
3. van Besouw, R.M., J.S. Brereton & D.M. Howard. 2008. Range of tuning for tones with and without vibrato. *Music Percept.* **26:** 145–156.
4. Shonle, J. & K. Horan. 1980. The pitch of vibrato tones. *J. Acoust. Soc. Am.* **67:** 246–252.
5. Iwamiya, S., K. Kosugi & O. Kitamura. 1984. Perceived principal pitch of FM-AM tones when the phase difference is in-phase and anti-phase. *J. Acoust. Soc. Jpn.* **5:** 59–69.
6. d'Alessandro, C. & M. Castellengo. 1994. The pitch of short-duration vibrato tones. *J. Acoust. Soc. Am.* **95:** 1617–1630.
7. Gardner, T.J. & M.O. Magnasco. 2005. Instantaneous frequency decomposition: an application to spectrally sparse sounds with fast frequency modulations. *J. Acoust. Soc. Am.* **117:** 2896–2903.