

Principal pitch of frequency-modulated tones with asymmetrical modulation waveform: A comparison of models

Pablo E. Etchemendy,^{a)} Manuel C. Eguia, and Bruno Mesz

Laboratorio de Acústica y Percepción Sonora, Universidad Nacional de Quilmes, R. S. Pena 352 Bernal, B1876BXD Buenos Aires, Argentina

(Received 29 January 2013; revised 23 December 2013; accepted 13 January 2014)

In this work, the overall perceived pitch (principal pitch) of pure tones modulated in frequency with an asymmetric waveform is studied. The dependence of the principal pitch on the degree of asymmetric modulation was obtained from a psychophysical experiment. The modulation waveform consisted of a flat portion of constant frequency and two linear segments forming a peak. Consistent with previous results, significant pitch shifts with respect to the time-averaged geometric mean were observed. The direction of the shifts was always toward the flat portion of the modulation. The results from the psychophysical experiment, along with those obtained from previously reported studies, were compared with the predictions of six models of pitch perception proposed in the literature. Even though no single model was able to predict accurately the perceived pitch for all experiments, there were two models that give robust predictions that are within the range of acceptable tuning of modulated tones for almost all the cases. Both models point to the existence of an underlying “stability sensitive” mechanism for the computation of pitch that gives more weight to the portion of the stimuli where the frequency is changing more slowly.

© 2014 Acoustical Society of America. [<http://dx.doi.org/10.1121/1.4863649>]

PACS number(s): 43.66.Hg, 43.66.Ba, 43.64.Bt [SAF]

Pages: 1344–1355

I. INTRODUCTION

Non-stationary stimuli with a periodically varying frequency (such as a frequency modulated signal) can evoke different pitch percepts depending on the rate and the amplitude of the variation. If the rate is low, our auditory system is able to track the frequency variation, and a pitch contour is perceived. As the rate increases, and if the frequency fluctuation is not too wide, a single steady pitch emerges, called the *principal pitch*. This fact is exploited in music performance as the vibrato technique, where a note is typically varied up to a semitone with a mean frequency of modulation between 4 and 8 Hz.¹ The principal pitch has enough stability to be measured with a precision of a few cents.^{2,3}

Compared to stationary stimuli, the number of studies that have examined the pitch assignment for non-stationary tones is small, and only a few numerical models for the pitch perception of these kind of stimuli have been proposed so far. Early studies were focused on modulated tones with parameters matching those of the vibrato present in opera singing and string instruments. Seashore¹ found that the principal pitch was located close to the arithmetic mean of the extremes of the modulation. In 1980, more carefully designed experiments were conducted by Shonle and Horan.⁴ They used sinusoidal carriers modulated by triangular or square waveforms as stimuli, varying the modulation depth and the frequencies of the carrier and modulation. They obtained pitch matches with unmodulated tones using a method of adjustment. They found that the average pitch

matches were closer to the geometric, rather than the arithmetic, mean between the frequency extremes.

Somewhat different results are obtained when the modulation waveform is not symmetrical or the amplitude is also modulated. In a subsidiary experiment, Shonle and Horan observed that when the modulation waveform had a trapezoidal shape, the average pitch matches were closer to the flat portion of the modulation. A similar result, displaying an “attraction” of the pitch matches by the flat portions, was obtained by Sundberg³ using synthesized vowels and a half-wave rectified sine as a modulation waveform. From these results Sundberg concluded that the principal pitch can be predicted by the linear time average of the instantaneous frequency. Therefore the principal pitch is biased toward the steady portions of the modulation, where the stimulus stays constant. However, the pitch matches obtained by Shonle and Horan were even more shifted toward the flat portions of the modulation than the linear time-average predictions. They concluded that the steady part of the modulation has somewhat more influence on the principal pitch than the average of the varying part.

Gockel *et al.*⁵ further investigated this pitch shift for asymmetric modulating functions in a series of experiments for a wide range of modulation (5–20 Hz) and carrier (0.5–8.0 kHz) frequencies and two initial modulation phases (0 and π). They used a two-interval two-alternative procedure to obtain pitch matches between a sinusoidal tone of 0.4 s of duration modulated by the asymmetric waveform and an unmodulated sinusoidal tone. The modulation waveform was a custom smooth function that contained a slowly varying portion [see Eq. (1) in Gockel *et al.*⁵]. In all cases, the pitch matches were shifted toward the slowly varying portion with respect to the time average of the instantaneous

^{a)}Author to whom correspondence should be addressed. Electronic mail: pe@lapseo.org

frequency. These results are consistent with the idea that for the estimation of the principal pitch, more weighting is given to portions of the stimulus where the frequency is changing more slowly. They thus proposed a model in which the principal pitch is obtained from a weighted time average of estimates of the period (weighted averaged period, WAP). The weighting is inversely related to the rate of change of the period, therefore the slowly varying portions of the modulation contribute with a larger weight to the final pitch estimate than the rest of the modulation cycle.

Iwamiya *et al.*^{2,6,7} studied stimuli that were simultaneously modulated in amplitude (AM) and frequency (FM) with triangular waveforms. They observed that when the modulations started with the same phase, the principal pitch was higher than the frequency mean and that when the AM and FM started with opposite phase, the pitch was lower than the predicted mean. These results are consistent with a weighted time average (WTA) model, where the weighting is related to the instantaneous amplitude. They proposed a power function of the instantaneous amplitude as weighting for the model (intensity-weighted average of instantaneous frequency, IWAIF). Based on a slightly different kind of stimulus (two sinusoids with closely spaced frequencies and different amplitudes), Feth⁸ found a similar result. To account for their results, they also used a WTA model based on the envelope of the function (envelope-weighted average of instantaneous frequency, EWAIF).

The effect of the starting phase of the modulation was explored by d'Alessandro and Castellengo,⁹ Gockel *et al.*,⁵ and van Besouw and Howard.¹⁰ While little or no influence was reported by Gockel *et al.*⁵ using modulated tones with less than or equal to two cycles, d'Alessandro and Castellengo⁹ observed an effect of phase on pitch for short duration modulated tones (up to two cycles of modulation). The authors postulated that the later parts of the tone have more weight in the overall pitch judgment and proposed a WTA memory model where the weighting is an exponential memory function that decays backward in time. Recently van Besouw and Howard¹⁰ studied the effect of phase in long duration FM tones. They found a small but statistically significant effect of the phase but in contradiction with the results of d'Alessandro and Castellengo. The authors argued that the difference is probably due to the presentation order of the tones that was not modified during the experiments.¹⁰

From these results, it can be conjectured that the perception of the principal pitch of modulated tones involves some sort of time average of the instantaneous frequency that could include the rate of change, the instantaneous amplitude, and some memory effects. However, the mechanisms underlying this temporal average are largely unknown.

More recently, another approach was put forward by Mesz and Eguia,¹¹ based on a nonlinear time-frequency representation proposed by Gardner and Magnasco.^{12,13} This model also computes the principal pitch as a WTA of the instantaneous frequency, but the weighting is obtained by a measure of "consensus" of the extracted instantaneous frequency between neighboring channels (where "channels" here are the set of discrete frequencies obtained from a short-time Fourier transform of the signal). There are

similarities between the weighting functions of the model of Gockel *et al.* and the model of Mesz and Eguia: When the rate of change of the instantaneous frequency is low, the consensus measure is high and the portion of the modulation contributes significantly to the overall pitch, and conversely when the instantaneous frequency varies rapidly, the consensus measure is low and the weighting function has a smaller value. Gockel *et al.* also proposed that there is a certain "sluggishness" in the pitch tracking of portions of sounds with fast frequency changes; this leads to the idea of a "stability-sensitive" perceptual weighting, biased in the direction of sound segments with slow frequency fluctuations. The measure of consensus proposed by Gardner and Magnasco¹² also exhibits this sensitivity because less stable sounds give rise to slightly different instantaneous frequency estimates across channels, thus lowering the consensus measure.

The purpose of this paper is to compare these models using sinusoidal tones modulated by a waveform that can be parametrically varied from symmetrical (triangular) to asymmetrical (trapezoidal). At one extreme, there is the original triangular vibrato tone studied by Shonle and Horan, where the principal pitch corresponds to the geometric mean. As the degree of asymmetry increases, the flat portion of the trapezoidal modulation is presumed to give more weight to the pitch estimation, which departs from the geometric mean. At the other extreme, the flat portion occupies the entire cycle, and there is a single frequency corresponding to the unmodulated signal.

Although asymmetric vibratos may sound unfamiliar to Western listeners, examples of similar frequency profiles can be found in South Indian classical (Carnatic) music.¹⁴ The pitch tracking of certain pitch inflections (gamakams) used in Carnatic music showed that the instantaneous frequency profiles can be approximated as sequences of "flat portions" and "peaks" that look similar to the trapezoidal modulations. Moreover, the shapes and heights of the peaks are highly variable even when the intonation is well preserved,¹⁴ a feature that suggests that the overall pitch of gamakams could also be dominated by the flat portions of the modulations.

The paper is organized as follows. In Sec. II, we review some of the pitch perception models presented in the literature, especially those that can be applied to non-stationary stimuli. Section III is devoted to the description of the experimental methods. Then the results are presented and analyzed in Sec. IV. A comparison of the predictions of the models described in Sec. II with the results presented in Sec. III and in previously reported experiments is discussed in Sec. V. Finally, Sec. VI provides a conclusion.

II. MODELS

In this section, we describe the pitch models that will be compared in this study. The first two models are based on the two main strategies for extracting pitch from stationary tones: Place-rate and temporal.¹⁵ For the place-rate representation, we calculate the excitation patterns derived from auditory filter banks¹⁶ and derive three different pitch estimators (see Sec. II A). For the temporal representation,

we calculate autocorrelation functions¹⁷ and derive two pitch estimators (see Sec. II B). The next three models (IWAIF/EWAIF, WAP, and consensus-based) were chosen because they were specifically proposed for predicting the pitch of non-stationary tones.

The choice of pitch perception models is not meant to be exhaustive but rather to provide a broad range of possible mechanisms for principal pitch extraction for non-stationary tones.

A. Excitation patterns

The first model to be considered is based on the long-term spectrum of the stimulus: The excitation patterns derived from the response of auditory filter banks as a function of their characteristic frequency (CF). These patterns correspond to the distribution of activity evoked along the basilar membrane.

From the place-rate theory point of view, changes in pitch are detected by shifts of the excitation patterns along the frequency axis. Because there is also a substantial shift in frequency of the peaks (place of maximum excitation) of these patterns with the intensity of the signal, this feature is often disregarded as an absolute indicator of pitch change. Instead changes in pitch are assumed to be detected by monitoring changes in level at one point or at multiple points on the excitation pattern. Zwicker¹⁸ proposed that when the change in the excitation level of the single point of the pattern that changes most exceeds 1 dB, a change in either loudness or pitch is detected. Some authors have noted that the low-frequency (apical) edge of the excitation pattern does not depend on sound intensity, and therefore they proposed that the changes in pitch are related to shifts of this low-frequency edge along the frequency axis.¹⁹

We follow the procedure described by Glasberg and Moore¹⁶ (with the correction suggested by the authors) to derive the auditory filter shapes and to compute the excitation patterns of the stimuli used in the experiments. From these excitation patterns, we derive three measures for pitch discriminability: (a) Zwicker criterion, testing whether two patterns differ by at least 1 dB; (b) frequency shifts of the low frequency slope; and (c) frequency shifts of the excitation pattern maxima.

B. Autocorrelation

Autocorrelation functions (ACF) have been proposed as predictors of pitch, either from the stimulus waveform^{17,20} or from the output of an auditory model.^{21,22} ACFs display several modes that correspond to the multiples of the period (or periods) present in the signal. A simple approach is to take the first large mode at nonzero lag as the cue to pitch,²⁰ although the combination with information obtained from the other modes gives more robust estimates.²²

We will take two estimates for pitch: (a) An average of the first ten peaks of the ACF of the signal and (b) the most salient pitch from a model proposed by Cariani²² based on the ACFs of the post-stimulus responses of an auditory nerve fiber (ANF) model. We selected this model because it also provides measures of pitch salience (the relative weight of a

pitch value when many virtual pitches are competing²³), and it was applied to non-stationary stimuli.

A short description of the model proposed by Cariani follows. First, the signal is processed by a simplified auditory periphery model that simulates cochlear and neural processes and splits it up into channels (corresponding to the ANFs with different characteristic frequencies and spontaneous rate). ACFs of the peristimulus signals are computed for each channel and summed across channels, weighted by ANF human SR and CF distribution, to produce a summary population time-interval distribution (PID). Next, an exponential-tapering weight that declines with interval duration with a time constant of 10 ms is applied to the PID to account for the lower limit of musical pitch²⁴ (~ 30 Hz) and the limited resolution of the model. Finally, a dense set of periodic sieves (25–1000 Hz in 2 Hz steps) is applied to the duration-weighted PID to measure the relative strength of interval patterns associated with different perceived pitches. The salience of a particular pitch is estimated by dividing the mean density (intervals/bin) of pitch-related intervals in sieve bins by the mean density of the whole distribution.

C. EWAIF and IWAIF models

The first WTA model was proposed by Feth⁸ to account for the discriminability of two-tone complexes and is based on the instantaneous frequency of the signal. In general, a finite real signal $x(t)$ can be represented as the real part of an analytic signal $m(t)$

$$x(t) = \text{Re}[m(t)] \quad (1)$$

with the analytic signal written as

$$m(t) = a(t)e^{-i\phi(t)}, \quad (2)$$

where the function $a(t)$ is the instantaneous envelope, and $\phi(t)$ the instantaneous phase. The instantaneous frequency is then defined as the time derivative of this phase

$$f_i(t) = \frac{1}{2\pi} \frac{d\phi(t)}{dt}. \quad (3)$$

The model proposed by Feth predicts the pitch as an EWAIF

$$f_{\text{EWAIF}} = \frac{\int a(t)f_i(t)dt}{\int a(t)dt}, \quad (4)$$

where f_{EWAIF} correspond to the frequency associated to the pitch predicted by the EWAIF model.

In later works, models based on the IWAIF have been proposed,^{6,25,26} but the general principle of the models remains the same.

D. WAP model

The second weighted averaged model aimed to predict the pitch of FM tones was proposed by Gockel *et al.*⁵ This model is based on short time estimates of the instantaneous waveform period instead of the instantaneous frequency.

For each cycle of the waveform i , the local period P_i is calculated as the time interval between two zero crossings. The weighting function W_i is related to the rate of change of this local period and is calculated as a function of the relative standard deviation (R_i) of the period over five consecutive cycles. The function proposed by the authors is

$$W_i(R_i) = 1/(1 + KR_i^{0.5}), \quad (5)$$

where K is a constant. The predicted frequency f_{WAP} is the inverse of the weighted average period

$$f_{\text{WAP}} = \frac{\sum W_i(R_i)}{\sum P_i W_i(R_i)}. \quad (6)$$

E. Consensus-based model

The third model was proposed by Mesz and Eguia.¹¹ As with the previous weighted average models, the principal pitch extraction consists of two stages: (a) The computation of the instantaneous frequency or period of the signal and (b) the derivation of a weighting function. While in the previous cases these two calculations were performed within the time domain, in this model, a nonlinear time-frequency representation, belonging to the reassignment class,¹² is used. This particular representation is intended to provide an accurate representation of rapidly changing sounds and is based on the notion of ‘‘consensus’’ that measures the consistency of the instantaneous frequency estimations along the frequency coordinate.

In the first stage of the model, the instantaneous frequency of the signal is derived from the reassigned frequencies computed with the method described by Fulop and Fitz.²⁷ The reassigned frequency or *channelized instantaneous frequency* is given by

$$\text{CIF}(x; t, \omega) = \frac{\partial}{\partial t} \arg(\text{STFT}_h(x; t, \omega)), \quad (7)$$

where $\text{STFT}_h(x; t, \omega)$ is the short-time Fourier transform of the signal $x(t)$, computed using the windowing function h as a function of time at a set of discrete frequencies ω , referred to as *channels*

$$\text{STFT}_h(x; t, \omega) = \int x(t + \tau)h(-\tau)e^{-i\omega\tau}d\tau. \quad (8)$$

For a given signal, $\text{CIF}(t, \omega)$ corresponds to the instantaneous frequency as a function of time for each channel ω . Then, the instantaneous frequency $F_i(t)$ of the whole signal is calculated as an average over channels of these frequencies, weighted by the power of the channel

$$F_i(t) = \frac{\sum_{\omega} \text{CIF}(x; t, \omega)X(t, \omega)^2}{\sum_{\omega} X(t, \omega)^2}, \quad (9)$$

where $X(t, \omega)$ is the modulus of the channel amplitude

$$X(t, \omega) = |\text{STFT}_h(x, t, \omega)|. \quad (10)$$

In the second stage, the weighting function is derived from the measure of consensus introduced by Gardner and Magnasco.¹² Consensus quantifies the similarity of the local instantaneous frequency estimates (CIF) in a group of neighboring channels. If these estimates are very similar the consensus is high. As it was suggested,²⁸ the mixed partial phase derivatives (MPD) of the STFT can be used as a local consensus measure. In effect, we have

$$\text{MPD}(t, \omega) = \frac{\partial^2}{\partial t \partial \omega} \arg(\text{STFT}_h(x; t, \omega)), \quad (11)$$

$$= \frac{\partial}{\partial \omega} \text{CIF}(x; t, \omega). \quad (12)$$

Hence, the modulus of the MPD also measures the rate of change of the CIF across channels. A small (respectively, big) MPD modulus can then be regarded as indicative of high (respectively, low) consensus.

The weighting function is constructed in such a way that a maximum value corresponds to maximum consensus and decays exponentially with decreasing consensus. This weighting function is also averaged proportionally to the energy of the channel

$$W(t; \gamma) = \frac{\sum_{\omega} C(t, \omega; \gamma)X(t, \omega)^2}{\sum_{\omega} X(t, \omega)^2}, \quad (13)$$

with γ a constant. The consensus matrix $C(t, \omega; \gamma)$ is expressed as a function of the mixed partial derivative of the channel amplitude

$$C(t, \omega; \gamma) = e^{-\frac{|\text{MPD}(t, \omega)|}{\gamma}}. \quad (14)$$

Note that for $\gamma \rightarrow 0$, this tends to emphasize the points where the consensus expressed as $\text{MPD}(t, \omega)$ is close to zero (high consensus) and to neglect the others, while for $\gamma \rightarrow \infty$, it gives equal relevance to all MPD values.

Finally, the predicted frequency for the consensus model (f_{con}) is given by

$$f_{\text{con}}(\gamma) = \frac{\sum_t F_i(t)W(t; \gamma)}{\sum_t W(t; \gamma)}. \quad (15)$$

III. EXPERIMENTS

A. Experiment 1

The aim of this experiment was twofold: (a) To measure the principal pitch of sinusoids modulated in frequency with a waveform the degree of asymmetry of which was varied parametrically and (b) to compare the predictions of the models presented in the previous section.

The modulation profiles were trapezoidal, each cycle consisting of a flat portion of constant frequency and two linear segments (on a log scale) forming a peak. The peaks

could be pointing toward the high frequencies (**u** profile) or toward the low frequencies (**n** profile). The asymmetry of the modulation is measured by the percentage of the cycle occupied by the flat portion (p_f). When $p_f = 0$, the modulation shape is triangular (symmetric), and there is no distinction between both cases, except for the starting phase (ψ).

1. Stimuli

The stimuli consisted of sinusoids modulated in frequency with the waveform described in the preceding text. The asymmetry of the modulation was used as a parameter. The chosen values of p_f were: 30%, 55%, and 75% for both profiles. Also, a symmetric profile was included ($p_f = 0\%$), which served as a control. This made a total of seven different stimuli.

All stimuli consisted of four cycles of modulation. The time-averaged geometric mean (f_c) was 1 kHz, and a modulation rate (f_{FM}) of 10 Hz was used. This modulation rate was chosen because preliminary listening tests showed that it is fast enough to avoid the tracking of the instantaneous frequency profile. The overall depth (measured from the maximum to the minimum frequency of the modulation profile) was $d_{FM} = 125$ cents. The starting phase (ψ) was varied randomly to wash out possible start or end effects. Finally, the amplitude was kept constant during the 400 ms presentation time of each stimulus, except for 5 ms raised-cosine onset and offset ramps. An example of the stimulus used is displayed in Fig. 1.

All stimuli were generated digitally with MATLAB at a sampling rate of 48 kHz with 16 bits of resolution, converted through a Focusrite Sapphire Pro 40 sound board, and presented diotically using previously calibrated Sennheiser HD 240 headphones. The nominal signal level was 65 dB sound pressure level (SPL).

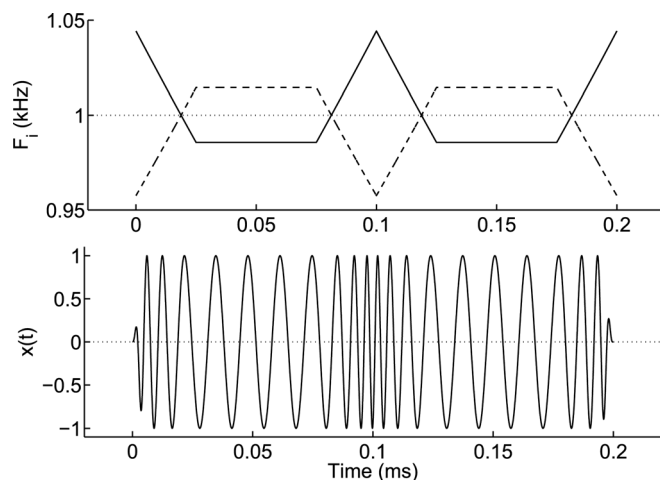


FIG. 1. The top panel shows the instantaneous frequency trajectories of the **n** shape modulation (dashed line) and **u** shape modulation (solid line) function. Both stimuli have a geometric mean (f_c) of 1 kHz, while the amplitude of modulation d_{FM} equals 100 cents, the modulation frequency f_{FM} is 10 Hz, and the starting phase of the modulation ψ is zero. The bottom panel shows the waveform of a sinusoid modulated by two cycles of the **u** function with 5 ms raised cosine onset and offset ramps. For clarity, a f_c of 200 Hz and a greatly exaggerated d_{FM} of 2000 cents are shown.

2. Subjects

Eight subjects participated in the experiment. All reported having normal hearing although this was not checked. Their ages ranged from 27 to 43 yr. Seven of them have a high degree of musical training with a degree in musical composition and with more than 15 yr of experience as players of wind instruments (MP, LN), string instruments (LS, EA), piano (HK, AA), and percussion (FN). One of them (LN) has experience as orchestral conductor. One of the authors (ME) also participated in the experiment.

3. Procedure

The experiment was divided into 14 blocks, two for each condition (modulation profile). Each block required 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 converge the pitch matches. The method adopted here is very similar to that used by Gockel *et al.*⁵

To check for possible order biases, we performed two blocks for each modulation. In one of the blocks, the modulated tone was presented in the first place, followed by the pure tone, after an inter-stimulus interval of 500 ms. In the other block, the presentation order was inverted. In both cases, the subject was asked to judge which tone had the higher pitch by pressing the key “1” (for selecting the first tone) or the key “0” (for selecting the second tone) on the keyboard. The stimuli pair could be repeated at will by pressing the space bar. We used a two-up two-down staircase procedure to adjust the frequency of the unmodulated tone with an adaptive step (s). After the first presentation, if the pitch of the pure tone was judged higher (respectively, lower), its frequency was adjusted with a negative (respectively, positive) step. Then the frequency of the pure tone was adjusted with the previous step unless the following “turning point” condition was met: The pure tone was judged “higher” (“lower”) in two consecutive presentations, and the step was negative (positive). In the turning point, the direction of the step is reversed. The initial value of s was 34.3 cents, and it was halved after each one of the first three turning points. The final step was small enough to allow from five to seven steps between turning points.

When the subject judged both pitches equal, he or she could inform this by pressing the “6” key. The frequency of the sinusoid was computed as a match if at least three turning points occurred. Then the procedure was repeated from the start until four pitch matches were obtained.

The initial frequency of the sinusoid was established in the following manner. For the first match, the starting frequency of the pure tone was 87 cents higher or lower, at random, than the central frequency of the vibrato. After the first match, the initial interval was 62 cents above or below the frequency of the last match. The direction of the starting interval was opposite with respect to the previous one.

The blocks were presented in random order except for the block with $p_f = 0\%$, which was presented always at the first place. The subjects were obliged to take a pause of

10 min after the seventh block to avoid possible effects of fatigue.

Before the beginning of the first block, subjects were familiarized with the procedure by making one match. A full test session lasted between 80 and 120 min.

B. Experiment 2

The aim of this experiment was to explore whether the effect of the asymmetry of the modulation could be changed by varying the central frequency. Following experiment 1, the degree of asymmetry (p_f) was varied parametrically, though adding a fifth p_f value.

1. Stimuli

The stimuli were similar to those of the previous experiment, except that the central frequency was 2 kHz and the chosen values of p_f were 0%, 24%, 30%, 55%, and 75% for both **u** and **n** profiles, giving a total of nine experimental conditions. All stimuli were generated and presented with the same equipment and the same features as in experiment 1.

2. Subjects

Four subjects participated of this experiment: Three (MP, HK, EA) who also participated in experiment 1 and one of the authors (BM).

3. Procedure

As in the previous experiments, each block consisted in obtaining four pitch matches between the modulated tone and an unmodulated sinusoid. In this case, only one block was employed for each type of modulation, making a total of nine blocks. The unmodulated tone was always presented in the first place.

The nine conditions were presented in random order, except for the block with $p_f = 0\%$, which was presented first. The subjects were obliged to take a pause of 15 min after the fourth block to avoid possible effects of fatigue. A full data session lasted between 40 and 90 min.

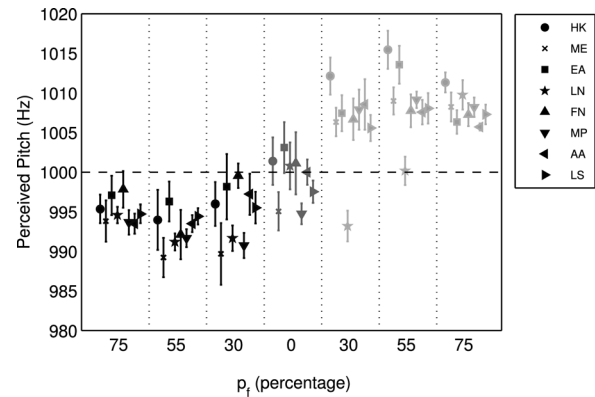


FIG. 2. Mean results and standard errors across trials for all eight subjects who participated in experiment 1. The geometric mean of stimulus, indicated with dashed line, was 1000 Hz with d_{FM} equal to 125 cents. Data are presented as a function of the type of modulation. Gray markers indicate **n** stimuli while black markers indicate the **u** stimuli. The frequencies were averaged over the two possible orders of presentation.

IV. RESULTS

A. Experiment 1

In Fig. 2, we display the means across trials for each subject, along with their standard errors. Means across subjects with the standard deviations of individual means can be found in Fig. 3 (where we also included the predictions of models, which will be discussed in Sec. V). The x axis indicates the degree of asymmetry of the modulation (p_f) and the orientation of the profile (**n** or **u**). The adjusted frequencies were averaged for the two possible orders of presentation, as this factor did not show a statistically significant effect (see following text).

Comparing individual differences, it is apparent that all subjects reported perceiving the principal pitch of **u** tones lower than the geometric mean, and all subjects but one (LN) reported perceiving the principal pitch of **n** tones higher than the geometric mean. The subject LN reported that he intentionally focused on the lower part of the modulation because he interpreted the modulated tones as trills, where the lower note is the base of the trill and is written in the score.

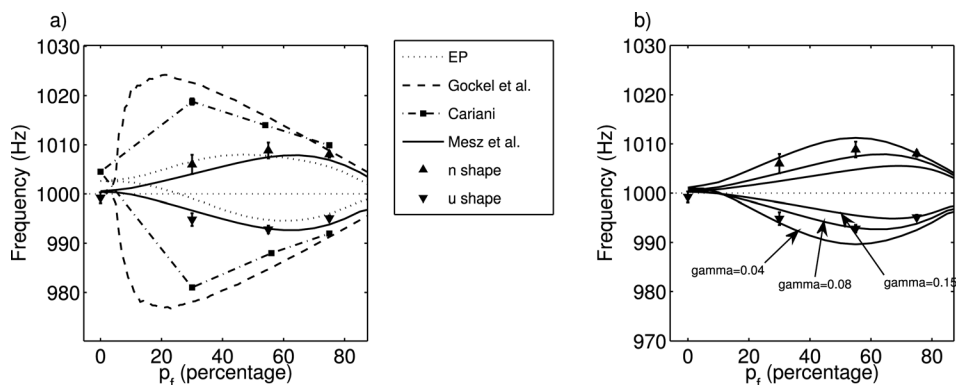


FIG. 3. Mean results across subjects for experiment 1, with the standard deviations of the individual means, are displayed with triangles (upward-pointing triangles for **n** stimuli, downward-pointing triangles for **u** stimuli). Left panel (a) allows the comparison between data and the predictions of the models described in the text. Right panel (b) allows the comparison between data and the predictions of the consensus model computed with three different values of its free parameter ($\gamma = 0.04, 0.08, \text{ and } 0.15$). The results are displayed as a function of the degree of asymmetry (p_f). All stimuli had a central frequency (f_c) of 1 kHz.

For means across subjects corresponding to **n** tones [Fig. 3(a), upward-pointing triangles], we observe a systematic increase of the adjustable frequency as p_f increases. The maximum deviation with respect to the geometric time average of the modulation profile (f_c) is 9 Hz (15 cents) and is reached when $p_f = 55\%$. A similar pattern is observed for data corresponding to **u** tones (downward-pointing triangles) for which the adjustable frequency decreases as p_f increases. The maximum deviation is 7.3 Hz (12.7 cents), also reached when $p_f = 55\%$.

A within-subjects analyses of variance (ANOVA) was conducted on the adjusted frequencies, with presentation order and type of stimuli as factors. The sphericity condition was met (Mauchly's test with $\chi^2(6) < 0.002$, $p > 0.12$ for both factor type of stimuli and for the interaction). There was a significant effect of the type of stimuli [$F(6, 42) = 42.4$; $p < 0.001$]. Neither the order of presentation nor the interaction showed a significant effect [$F(1, 7) = 2.85$ with $p = 0.14$ and $F(6, 42) = 1.35$ with $p = 0.26$, respectively].

We also carried out two sets of comparisons via paired t -tests to evaluate: (a) If there is a shift in the perceived pitch with respect to the geometric mean due to the asymmetry of the modulation and (b) if these shifts change for different degrees of asymmetry. The comparisons are listed in Table I. The first set of comparisons was performed between the results obtained for the stimulus with symmetric modulation and all the other stimuli with asymmetric modulation. For the case of the **u** type of modulation, we wanted to test whether the perceived pitch is higher than its geometric mean or not, while for the **n** type of modulation, we wanted to test if it is lower. Hence, we are only interested in one side of the statistical distribution and the tests are one-tailed. This set of comparisons gave significant results, except for the stimulus with **n** profile and 30% of flat portion. The second set of comparisons was performed between stimuli with

TABLE I. Statistical tests for pairs of stimuli for experiment 1. The tests consisted of paired t -tests. The experimentwise level of significance used was 0.05, and it was adjusted with a Bonferroni correction for each individual comparison. The comparisons against the symmetric stimuli were one-tailed; the remaining comparisons were two-tailed. Columns A and B refers to the two stimuli compared with the subcolumns indicating the values of their parameters. The comparisons reveals the effect of the parameter controlling the degree of asymmetry of the modulation function (p_f) plus the effect of the modulation profile. The last column shows the statistical value of each comparison.

A		B		p
p_f	Shape	p_f	Shape	
0	Symm	30	u	0.0007*
0	Symm	55	u	<0.0001*
0	Symm	75	u	0.0007*
0	Symm	30	n	0.0106
0	Symm	55	n	0.0005*
0	Symm	75	n	0.0001*
55	u	30	u	0.062
55	u	75	u	0.018
55	n	30	n	0.37
55	n	75	n	0.64

the same type of asymmetric modulation (**u** vs **u** and **n** vs **n**) to test whether significant differences exist among the different degrees of asymmetry explored. In this case, the tests were two-tailed because we did not make an *a priori* hypothesis about which stimulus was expected to have a pitch with a larger departure from the geometric mean. This set of comparisons gave no significant difference between pairs of stimuli. In all cases, the adjusted frequencies were averaged between orders of presentation for each subject. The experiment-wise level of significance was 5%. The level of significance of each comparison was adjusted using the Bonferroni correction.

B. Experiment 2

In Fig. 4, we display the means across trials for each subject, along with their standard errors, for the different modulation profiles indicated on the x axis. Means across subjects of the individual means along with their standard deviations can be found in Fig. 5.

As in experiment 1, data corresponding to tones with **n** profiles (upward-pointing triangles) show a systematic increase in the adjustable frequency as the percentage of the flat portion of the modulating function increases. The opposite shift is observed in data corresponding to **u** tones (downward-pointing triangles). A slight asymmetry is observed when comparing the shifts of the **n** and **u** profiles.

A within-subjects ANOVA was conducted on the adjusted frequencies with type of stimuli as factor. The analysis gave a significant effect of this factor [$F(8, 24) = 61.17$, $p < 0.001$].

V. DISCUSSION

The results of experiments 1 and 2 show that the pitch of a sinusoidal carrier modulated in frequency with an asymmetric profile deviates from the geometric mean. The direction of these pitch shifts is consistent with the hypothesis described in the Introduction: That a “stability-sensitive” perceptual weighting mechanism is involved in the computation of pitch, giving a result biased toward the pitch of sound

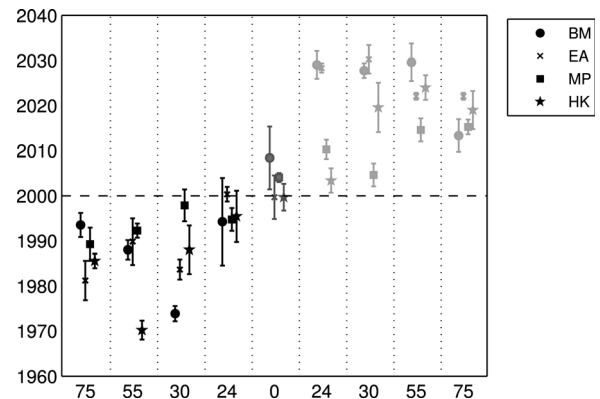


FIG. 4. Mean results and standard errors across trials for all four subjects who participated in experiment 2. The geometric mean of stimulus, indicated with dashed line, was 2000 Hz with d_{FM} was 125 cents. Gray markers indicate **n** stimuli while black markers indicate the **u** stimuli.

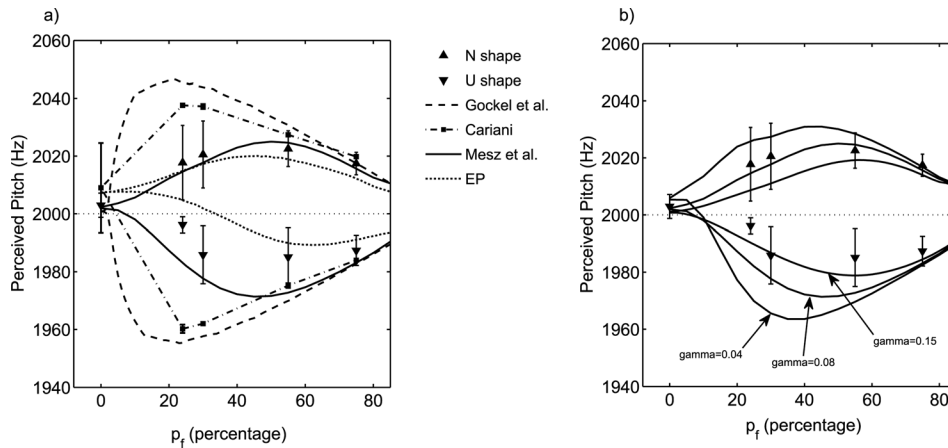


FIG. 5. Mean results across subjects for experiment 2 with the standard deviations of the individual means displayed as triangles (upward-pointing triangles for **n** stimuli, downward-pointing triangles for **u** stimuli). Left panel (a) allows the comparison between data and the predictions of the models described in the text. Right panel (b) allows the comparison between data and the predictions of the consensus model computed with three different values of its free parameter ($\gamma = 0.04, 0.08, \text{ and } 0.15$). The results are displayed as a function of the degree of asymmetry (p_f). All stimuli had a central frequency (f_c) of 2 kHz.

segments with slow frequency fluctuations, which in this study corresponds to the steady portion of the modulation.

These results are consistent with the findings of Gockel *et al.*⁵ because the corresponding degree of asymmetry required for a trapezoidal stimulus to approximate the modulation profile used by the authors is ($p_f = 65\%$). This is a value that falls within the range where we found a statistically significant deviation in pitch from the geometrical mean.

A. Comparison of pitch models

We now turn to the comparison of the prediction of the models described in Sec. II with the results obtained in our experiment.

We calculated the excitation patterns for all our stimuli using the method described in Sec. II A and compared the profiles obtained for **u** and **n** type stimuli having same geometric mean to determine whether or not differences exist between them. In Fig. 6, we display excitation patterns for **u** type (black) and **n** type (gray) stimuli with SPL 60 dB, $p_f = 75\%$ and $f_c = 1000$ Hz (solid line). The long-term power

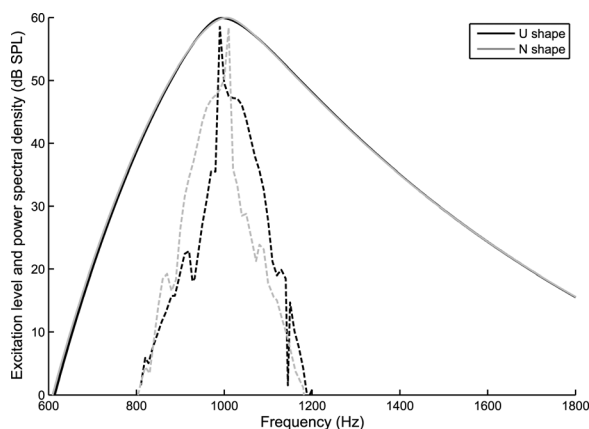


FIG. 6. The black and gray dashed lines show long-term spectra for the **u** and **n** stimuli, respectively. The smooth curves in solid line show corresponding excitation patterns (same color code for the modulation shape). The stimuli displayed had SPL 60 dB, $p_f = 75\%$ and $f_c = 1000$ Hz.

spectra of these stimuli are also displayed as a reference (dotted line). From the figure it is apparent that the differences of level between the excitation patterns are much more subtle than the difference between the spectra. In fact, the excitation pattern difference does not satisfy the Zwicker criterion, i.e., the two curves are less than 1 dB apart at any frequency, within the range between 20 and 60 dB, for all stimuli used in our experiment. The low-frequency slope criterion is not applicable for these excitation patterns because the slope is not uniform, causing a crossing between patterns about 10 dB below the peak. The only observable difference that can be extracted from the excitation patterns is the position of the maximum. In all cases, the frequency of the peak for the pattern corresponding to the **n** profile stimulus was higher than the corresponding frequency for the **u** profile stimulus. In Figs. 3 and 5, we display in dotted lines the frequencies of the EP peaks for comparison with the frequencies of the pitch matched data. The differences between these values are larger for the **u** profiles due to the fact that the peak shifts are not symmetrical with respect to the central frequency.

We then compared the shifts of the first ten peaks of the ACFs of the signal for the **u** and **n** shape profile stimuli to the multiples of the mean period $1/f_c$. In all cases (all frequencies and shapes), the shifts are in the same direction as the results of our experiments and display a similar dependence on the parameter p_f with an extreme in the region between $p_f = 40\%$ and $p_f = 60\%$. The magnitude of the shift depends on the number of the peak for all possible values of p_f . A mean for the first five peaks gives a maximum shift of 2.6 cents while the same for the first ten peaks gives 7 cents, too small to explain the experimental data.

The predicted pitch values obtained from the model proposed by Cariani²² and described in Sec. II B are displayed in Figs. 3 and 5 as dashed-dotted lines for comparison with the pitch matched data. For all cases with asymmetrical modulation, the model predicts frequency shifts from the geometric mean in the same direction as those observed in our experiment. The magnitude of these shifts is higher than the average observed pitch shifts, notably for modulation

profiles with $p_f = 30\%$. For these modulation profiles, the difference between the model predictions and the mean value of the experimental results is greater than 20 cents.

Models based on the EWAIF or IWAIF (see Sec. II C) predict pitch values that in all cases are too close to the geometric mean to account for the observed shifts and are not displayed in the figures.

Then we computed the predicted frequencies, f_{WAP} , of the WAP model (see Sec. II D) using Eq. (6) with a parameter value of $K = 600$ (the same value used by Gockel *et al.* for adjusting the results of their experiments). As was the case with the model proposed by Cariani, the predicted shifts were always greater in absolute value than the average observed pitch shifts, particularly for $p_f = 30\%$. The predictions of the WAP model are displayed in Figs. 3 and 5 with a dashed line. In both cases, these differences occur for the **u** modulation profiles with $p_p = 30\%$. Note also that the functional form of f_{WAP} vs p_f has a maximum departure from the geometric mean near $p_f = 20\%$, while the maximum departure for the perceived pitch occurs always for values of p_f greater than 30%. Changing the value of the parameter K modifies the magnitude of the predicted shifts but the functional form is not substantially modified.

Finally, we analyzed the predictions of the consensus model. Because we did not fit the experimental data using the free parameter γ , we display in Figs. 3(b) and 5(b), the predictions of the model using three different γ values: 0.04, 0.08, and 0.15. For experiment 1, the predictions (with $\gamma = 0.08$) differ from the experimental mean data by at most 2 Hz (Fig. 3), which corresponds to a departure of 3.4 cents. The maximum departure also occurs for $p_f = 30\%$. For experiment 2, the differences are less than 3 Hz (Fig. 5), which corresponds to 2.6 cents, for the **n** modulation profiles. In this case, the difference between the predicted and observed mean values are greater for the **u** profiles, reaching 13 Hz (11 cents) for $p_f = 24\%$. Except for this last case, the differences are all within the RAT (range of acceptable tuning) for vibrato tones,²⁹ which goes up to 10 cents, so that a vibrato with carrier frequency of up to 10 cents flatter than the fundamental frequency of a tone without vibrato is still perceived as being in tune. The functional form of f_{con} vs p_f has a maximum departure from the geometric mean between $p_f = 50\%$ and $p_f = 70\%$ for experiment 1 and for a wide range of γ values, which agrees with the maximum departure for the perceived pitch that occurs for values $p_f = 55\%$ and $p_f = 75\%$. For the case of experiment 2, there is also an agreement in the range of p_f values of the maximum departure from the geometric mean. Indeed, the peak value of f_{con} is located between $p_f = 40\%$ and $p_f = 60\%$, while for the perceived pitch the maximum value occurs between $p_f = 30\%$ and $p_f = 55\%$.

B. Survey of previous results

We also examined the data in the existing vibrato literature and compared the reported results with the predictions of the models, with exception of the EWAIF and IWAIF models, given that all their predictions are too close to the geometric mean.

Shonle and Horan⁴ conducted a subsidiary experiment with trapezoidal frequency profiles having $p_f = 50\%$, a nominal modulation depth of 100 cents and a modulating rate of 6 Hz. The design of the experiment was different from ours with the frequency of the sine wave carrier being dependent on the modulation profile. For their “flat bottom” vibratos (corresponding to our **u** profiles), they used a carrier with f_c equal to 433.5 Hz and obtained a pitch of 432.9 ± 1.4 Hz. For their “flat top” vibratos (**n** profiles), the corresponding values were 446.4 Hz for the carrier and 448 ± 1.3 Hz for the perceived pitch. The mean values of the shifts with respect to the geometric mean were +6 and -2.5 cents, respectively.

We first analyzed the excitation patterns for these stimuli. Considering the **n** profile, the peak of the excitation pattern was located at 448.6 Hz (shift of +8 cents). For the **u** profile, the peak was located at 433 Hz (shift of -2 cents). For both values of f_c , the low-frequency slope between opposite profiles showed a separation lower than 0.2 dB. The crossing of the EP curves occurs at levels that are 5 dB below the peak.

The shifts for the first ten peaks of the ACFs of the signal show a pattern similar to that described in the previous section: They are in the same direction as the experimental data, and their magnitudes increase linearly with the number of the peak. Taking the mean for the first ten peaks gives maximum absolute shifts of 3 cents, too small compared with the experimental results.

Finally, the prediction given by the model of consensus falls inside the data range, while the WAP and the Cariani models predict shifts higher than 14 cents in absolute value, thus overestimating the values reported in the experiments.

In Fig. 7 we display the results reported by Iwamiya *et al.*⁶ The upper row shows the data corresponding to stimuli where the size of the amplitude modulation was fixed at $d_{\text{AM}} = 1$. For those experiments, the size of the frequency modulation varied from 0 (AM-only) up to 100 cents (mixed AM-FM). The lower row shows data for stimuli with a fixed size of frequency modulation ($d_{\text{FM}} = 100$ cents) and values of d_{AM} between 0 (FM-only) and 1 (mixed). Each column displays stimuli with a different central frequency (first column: $f_c = 440$ Hz; second column: $f_c = 880$ Hz; third column: $f_c = 1500$ Hz). In all cases, the authors used a sinusoidal carrier modulated with a triangular wave both for the amplitude and the frequency. Upward-pointing triangles correspond to the “in-phase” condition between both modulations while downward-pointing triangles correspond to the “anti-phase” condition.

The data show a clear pattern with shifts increasing as d_{AM} goes from 0 to 1 (upper row) and as d_{FM} goes from 0 to 100 cents (lower row). The shifts from the central frequency are positive when AM and FM are in phase, and negative when they are in anti-phase with maximum values between 5 cents ($f_c = 880$ Hz, anti-phase condition) and 13 cents ($f_c = 440$ Hz, in-phase condition). This behavior is accounted for by the different models we tried with discrepancies in the magnitude of shifts. The WAP model (dashed line) gives the closest prediction for most of the data points while the remaining models overestimate the shift by a factor of 2–4.

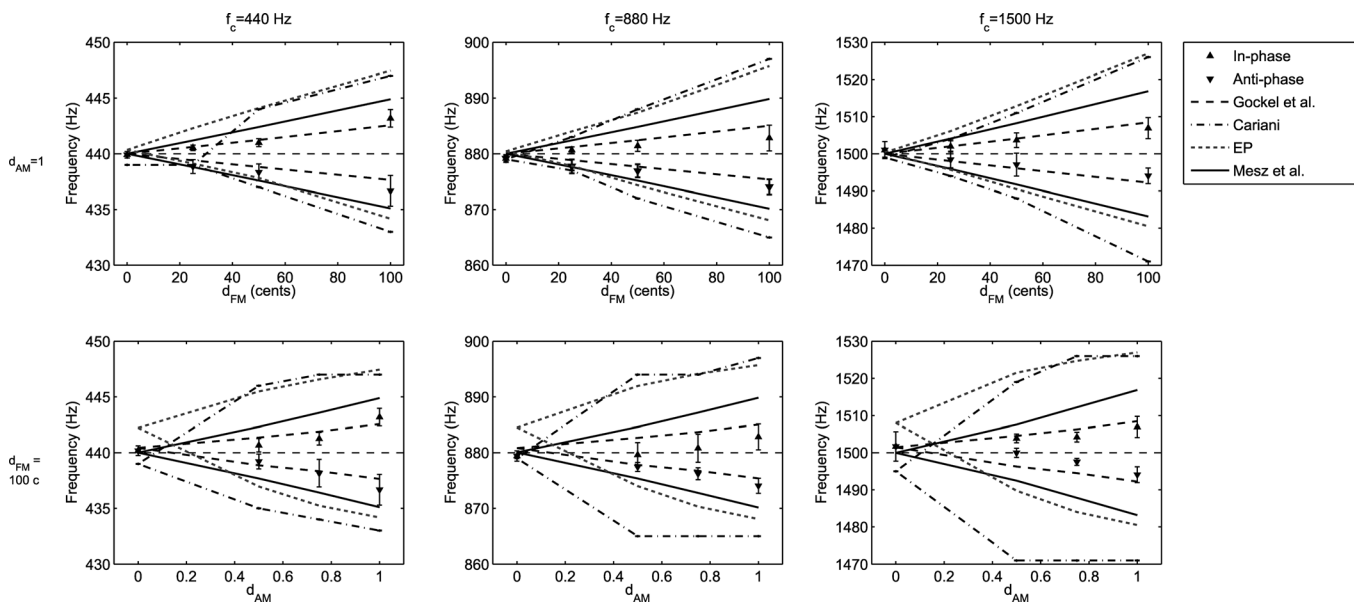


FIG. 7. Comparison of the data of Iwamiya *et al.* with the predictions of several models. The experimental data are shown with upward-pointing triangles (“in-phase” condition between modulations) and with downward-pointing triangles (“anti-phase” condition). Upper row indicates data of stimuli with the size of the amplitude modulation fixed at $d_{AM} = 1$ for which the size of the FM was parametrically varied. Lower row indicates data of stimuli with a fixed size of frequency modulation ($d_{FM} = 100$ cents), for which the size of the AM was varied between 0 and 1. The three columns of plots display stimuli with a different central frequency ($f_c = 440, 880,$ and 1500 Hz, respectively).

The last work we are going to review is a series of experiments made by Gockel *et al.*⁵ with asymmetrical frequency modulated sinusoids (see Sec. I). Results are displayed in Fig. 8; each plot corresponds to a different combination between central frequency (f_c) and modulation rate (d_{FM}). The perceived pitch is plotted against the modulation frequency (f_{FM}). Upward-pointing triangles correspond to the **n** modulation profile ($\cap\cap$ in terms of the authors) while downward-pointing triangles correspond to the **u** modulation profile ($\cup\cup$, respectively); black markers indicate “starting phase 1” ($\psi = 0$, the modulation starts at the extreme of the fast frequency excursion) while white markers indicates “starting phase 2” ($\psi = \pi$, the modulation starts halfway through the plateau in instantaneous frequency).

The shifts are mostly coincident with the predictions of the WAP model (dashed line), the values of which are contained inside the error bars or between the pitches corresponding to opposite starting phases. The model of

consensus (full black line) gives a good agreement also, although it shows a growth of the magnitude of the shifts with the modulation frequency, which seems to saturate for $d_{FM} > 10$ Hz. This departure from the data is shown clearly for $f_c = 500$ Hz and $d_{FM} = 8\%$. The model of Cariani (dashed-dotted line) gives a good agreement except for some departures when $f_{FM} = 20$ Hz. For the excitation patterns, their peaks (dotted line) give a slight sub-estimation of the shifts. The Zwicker criterion shows differences that are smaller than 1 dB. Moreover, all pairs of patterns show an intersection at intensities between 43 and 47 dB, making the criterion not applicable.

For the autocorrelation functions, the results obtained are similar to those of the literature cited before. The first ten peaks of ACF are shifted with respect to the position expected for the carrier wave ($1/f_c$). The shifts have the correct direction relative to the experimental data and show a similar dependence with respect to the modulation rate the magnitude of which depends on the number of the peaks in a

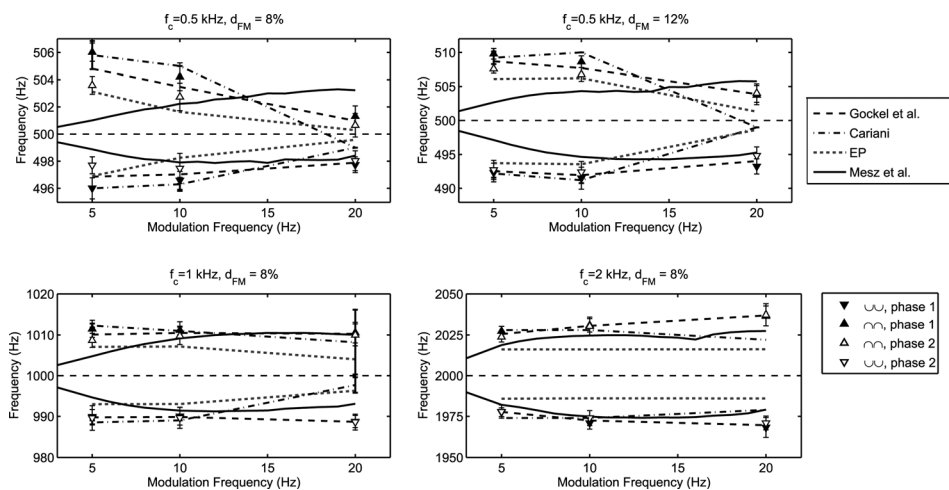


FIG. 8. Comparison of the data of Gockel *et al.* with the predictions of several models. Each plot shows data for a different combination of central frequency (f_c) and depth of modulation (d_{FM}). The experimental data are shown with downward-pointing triangles ($\cup\cup$ stimuli) and upward-pointing triangles ($\cap\cap$ stimuli). The starting phase is indicated with the color of the markers. The results are displayed as a function of the rate of the modulation (f_{FM}).

linear monotone-increasing fashion. However, in all cases, the shifts underestimated the experimental data.

A final point of discussion concerns the biological plausibility of the models. With the possible exception of that proposed by Cariani, the models are not formulated in terms of physiologically meaningful parameters. Hence we can only evaluate the models in terms of the constraints imposed by the available physiological and psychoacoustical data.

Concerning the model based on the excitation patterns, it is unlikely that the EP profiles corresponding to stimuli with opposite modulation shapes (e. g., **u** and **n**) could be discriminated. As it was noted before, changes in the EP profiles that are much less than 1 dB are not detectable.¹⁸ This was the case for the vast majority of the stimulus studied here. On the other hand, the difference between the EP maxima could not be detected because the level differences near the maxima are in all cases less than 0.25 dB.

With respect to autocorrelation-based models, the biological plausibility is still a matter of debate because of the requirement of long internal delay lines.^{30,31} In addition to this, the simplest models that calculate the ACF from the waveform require a high time resolution (up to hundreds of kHz) to achieve the accuracy needed for the determination of principal pitch in non-stationary tones. When the ACF is obtained summing the output of an auditory periphery model across frequency, this last requirement does not hold. In fact, Cariani and Delgutte have shown that pitch for non-stationary stimuli can be tracked from all-order histograms of auditory nerve fiber responses.³² All-order histograms are equivalent to the ACF and can be applied to a population of neurons. However, the underlying mechanisms for such processing are currently unknown.

As the WAP model is based on the analysis of the waveform, its major limitation also comes from the high precision required for determination of the period. If the period is estimated from the times between consecutive maxima, that is the most physiologically plausible situation, the time resolution needed largely surpasses the temporal precision in phase locking for a single channel.³³ However, it is possible that the auditory system have developed different strategies for enhancing the temporal precision comparing estimates coming from different channels.³⁴

The consensus-based model relies on the possibility of integrating temporal and rate-place mechanisms in the course of auditory neural processing. The integration in a “spatiotemporal” representation allows the extraction of a more reliable percept of pitch exploiting the place-rate and temporal coding redundancy^{35,36} and admits more realistic decoding mechanisms (without need of delay lines), such as across-channel coincidence detection. However, recently Carlyon *et al.*³⁷ have shown that straightforward pitch estimates obtained from this representation are strongly influenced by the overall level. Hence, the means whereby the pitch could be extracted under this approach are far from being understood.

Finally, it is possible that WAP and consensus-based models are describing the same underlying phenomenon (namely, the stability sensitive computation of pitch for non-stationary stimuli). Consensus performs essentially the same

role as the reciprocal of the rate of change of the instantaneous period. However, the consensus measure allows an additional interpretation. In fact, high consensus is attained when neighboring channels have very similar instantaneous frequency estimates; it can be indicative then of the reliability of the frequency estimation.¹³ Therefore the sound segments with the most reliable frequency determination (in this particular sense) are the most favored within the consensus-based model in their contribution to the overall pitch. Thus this model assumes that the overall pitch estimation is more based on the consistency of the temporal representation in groups of channels than on a fine frequency resolution from a single channel.

VI. SUMMARY AND CONCLUSIONS

In this study, we obtained principal pitch measures for vibrato tones (modulated sinusoids) varying the degree of asymmetry of the modulation. Consistently with previous results, we observed significant pitch shifts with respect to the geometric mean for vibratos modulated with waveforms with a degree of asymmetry greater than 30%. All significant shifts were in the direction of the flat portion of the modulation (upward in frequency for the **n** profiles and downward for the **u** profiles).

The predictions from six different pitch perception models were compared with the results obtained in our experiment and with previously reported data. The six models of choice were: (1) Pitch shifts derived from excitation pattern calculations following the method of Glasberg and Moore;¹⁶ (2) an average of the first peaks of the autocorrelation of the signal; (3) the autocorrelation model proposed by Cariani;²² (4) a simple weighted time average model (EWAIF/IWAIF);²⁵ (5) the model proposed by Gockel *et al.*⁵ (WAP); and (6) a consensus-based model proposed by Mesz and Eguia.¹¹ None of the above models accounted for the published data reviewed. However, models 1, 2, 3, 5, and 6 correctly predict the directions of the pitch shifts when these were significant. One major drawback of the model based on the excitation pattern (EP) calculations is that it relies on the position of the EP peak, a feature that is also dependent on sound level. The predictions of the model 2 also suffer a lack of robustness in the sense that they depend on the number of the peaks of the ACF that are taken into account. The three remaining models perform differently depending on the experimental data. The model proposed by Cariani generally overestimates the pitch shifts, giving the more accurate predictions for the experiments reported by Gockel *et al.* The WAP model gives the best performance for the data reported by Iwamiya and those of Gockel *et al.* Finally the consensus-based model gives accurate enough predictions (within the range of acceptable tuning) for all experiments except for some cases of the data reported by Gockel *et al.* Hence, models 5 and 6 appear to be the most accurate.

Because we did not adjust the parameters of the models, the comparison cannot be based on the accuracy of their predictions only. For that reason, we studied the functional dependence of the predicted pitch with the degree of

asymmetry, which is non-monotonic. The functional form that more resembles the behavior of the experimental results is that of the consensus-based model 6. However, further research is necessary to decide between these two models, possibly generating stimuli where the rate of change of the frequency and the consensus can be varied independently.

ACKNOWLEDGMENTS

We would like to thank Peter Cariani and Marcelo Magnasco for useful discussions. This work was partially funded by CONICET and UNQ.

- ¹C. Seashore, "The natural history of the vibrato," *Proc. Natl. Acad. Sci. U.S.A.* **17**, 623–626 (1931).
- ²S. Iwamiya, K. Kosugi, and O. Kitamura, "Perceived principal pitch of vibrato tones," *J. Acoust. Soc. Jpn. (E)* **4**, 73–82 (1983).
- ³J. Sundberg, "Effects of the vibrato and the 'singing formant' on pitch," *Musica Slovaca* **5**, 5–17 (1978).
- ⁴J. I. Shonle and K. E. Horan, "The pitch of vibrato tones," *J. Acoust. Soc. Am.* **67**, 246–252 (1980).
- ⁵H. Gockel, B. C. J. Moore, and R. P. Carlyon, "Influence of rate of change of frequency on the overall pitch of frequency-modulated tones," *J. Acoust. Soc. Am.* **109**, 701–712 (2001).
- ⁶S. Iwamiya, S. Nishikawa, and O. Kitamura, "Perceived principal pitch of fm-am tones when the phase difference between frequency modulation and amplitude modulation is in-phase and anti-phase," *J. Acoust. Soc. Jpn. (E)* **5**, 59–69 (1984).
- ⁷S. Iwamiya and K. Fujiwara, "Perceived principal pitch of fm-am tones as a function of the phase difference between frequency modulation and amplitude modulation," *J. Acoust. Soc. Jpn. (E)* **6**, 193–202 (1985).
- ⁸L. Feth, "Frequency discrimination of complex periodic tones," *Atten., Percept. Psychophys.* **15**, 375–378 (1974).
- ⁹C. d'Alessandro and M. Castellengo, "The pitch of short-duration vibrato tones," *J. Acoust. Soc. Am.* **95**, 1617–1630 (1994).
- ¹⁰R. M. V. Besouw and J. Brereton, "Effects of carrier and phase on the pitch of long-duration vibrato tones," *Musicae Sci.* **XIII**, 139–161 (2009).
- ¹¹B. A. Mesz and M. C. Eguía, "The pitch of vibrato tones: A model based on instantaneous frequency decomposition," *Ann. NY Acad. Sci.* **1169**, 126–130 (2009).
- ¹²T. J. Gardner and M. O. Magnasco, "Sparse time-frequency representations," *Pro. Natl. Acad. Sci. U.S.A.* **103**, 6094–6099 (2006).
- ¹³T. J. Gardner and M. O. Magnasco, "Instantaneous frequency decomposition: An application to spectrally sparse sounds with fast frequency modulations," *J. Acoust. Soc. Am.* **117**, 2896–2903 (2005).
- ¹⁴A. Krishnaswamy, "Pitch measurements versus perception of south indian classical music," in *Proceedings of the Stockholm Music Acoustics Conference (SMAC-03)*, Stockholm, Sweden, August 6–9 (2003).
- ¹⁵A. de Cheveigné, "Pitch perception," in *Oxford Handbook of Auditory Science: Auditory Perception*, edited by C. J. Plack (Springer, New York, 2010), pp. 71–104.
- ¹⁶B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.* **47**, 103–138 (1990).
- ¹⁷J. Licklider, "A duplex theory of pitch perception," *Cell. Mol. Life Sci.* **7**, 128–134 (1951).
- ¹⁸E. Zwicker, "Masking and psychological excitation as consequences of the ear's frequency analysis," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (Sijthoff, Leiden, Netherlands, 1970), p. 474.
- ¹⁹J. Zwillocki and M. Nguyen, "Place code for pitch: A necessary revision," *Acta Otolaryngol.* **119**, 140–145 (1999).
- ²⁰W. A. Yost, R. Patterson, and S. Sheft, "A time domain description for the pitch strength of iterated rippled noise," *J. Acoust. Soc. Am.* **99**, 1066–1078 (1996).
- ²¹R. Meddis and L. O'Mard, "A unitary model of pitch perception," *J. Acoust. Soc. Am.* **102**, 1811–1820 (1997).
- ²²P. Cariani, "A temporal model for pitch multiplicity and tonal consonance," in *Proceedings of the Eighth International Conference on Music Perception and Cognition (ICMPC8)*, Evanston, IL (2004), pp. 310–314.
- ²³E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Am.* **71**, 679–688 (1982).
- ²⁴D. Pressnitzer, R. D. Petterson, and K. Krumbholz, "The lower limit of melodic pitch," *J. Acoust. Soc. Am.* **109**, 2074–2084 (2001).
- ²⁵J. N. Anantharaman, A. K. Krishnamurthy, and L. L. Feth, "Intensity-weighted average of instantaneous frequency as a model for frequency discrimination," *J. Acoust. Soc. Am.* **94**, 723–729 (1993).
- ²⁶H. Dai, "On the pitch of two-tone complexes," *J. Acoust. Soc. Am.* **94**, 730–734 (1993).
- ²⁷S. A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *J. Acoust. Soc. Am.* **119**, 360–371 (2006).
- ²⁸D. J. Nelson, "Instantaneous higher order phase derivatives," *Digit. Signal Process.* **12**, 416–428 (2002).
- ²⁹R. M. V. Besouw, J. S. Brereton, and D. M. Howard, "Range of tuning for tones with and without vibrato," *Music Percept.* **26**, 145–156 (2008).
- ³⁰A. de Cheveigné and D. Pressnitzer, "The case of the missing delay lines: Synthetic delays obtained by cross-channel phase interaction," *J. Acoust. Soc. Am.* **119**, 3908–3918 (2006).
- ³¹R. Meddis and L. P. O'Mard, "Virtual pitch in a computational physiological model," *J. Acoust. Soc. Am.* **120**, 3861–3869 (2006).
- ³²P. Cariani and B. Delgutte, "Neural correlates of the pitch of complex tones. I. Pitch and pitch salience," *J. Neurophysiol.* **76**, 1698–1716 (1996).
- ³³C. Micheyl, B. C. J. Moore, and R. P. Carlyon, "The role of excitation-pattern cues and temporal cues in the frequency and modulation-rate discrimination of amplitude-modulated tones," *J. Acoust. Soc. Am.* **104**, 1039–1050 (1998).
- ³⁴B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. (Emerald, Bradford, UK, 2012), pp. 232–236.
- ³⁵A. J. Oxenham, J. G. W. Bernstein, and H. Penagos, "Correct tonotopic representation is necessary for complex pitch perception," *Proc. Natl. Acad. Sci. U.S.A.* **101**, 1421–1425 (2004).
- ³⁶L. Cedolin and B. Delgutte, "Spatiotemporal representation of the pitch of harmonic complex tones in the auditory nerve," *J. Neurosci.* **30**, 12712–12724 (2010).
- ³⁷R. P. Carlyon, C. J. Long, and C. Micheyl, "Across-channel timing differences as a potential code for the frequency of pure tones," *J. Assoc. Res. Otolaryngol.* **13**, 159–171 (2011).