ENEL 697 Term Project

Hough-Radon Transform for Detection of Vibrato

By Daniel Adler
Submitted to Dr. R.M. Rangayyan

11th April 2007

1 Introduction

We have developed a MATLAB application that detects instances of vocal vibrato from input digital audio files. The program performs vibrato detection using digital image processing and analysis methods in two dimensional time-frequency plane representations of the audio files. Detection of vibrato entails ascertaining values for a pre-defined collection of parameters that we deem appropriate to describe the physiologic phenomenon.

We begin with some basic background on vibrato and vocal formants. Knowledge of these two aspects of the voice is required to understand our choice of algorithms. We also give an introduction to time-frequency analysis using the spectrogram. This is followed by theoretical and implementation-specific details of the program's components. We show results of the program on both representative synthetic data and on actual data collected from opera singers.

1.1 Vibrato

Vocal coach David Jones defines vocal vibrato as a "slight variation of pitch resulting from the free oscillation of the vocal cords" [1]. It is detected audibly when listening to trained opera singers as a cyclic variation (or modulation) in pitch about the fundamental frequency of a note with time. Vibrato is also manifested as a modulation of upper harmonic frequencies (called "partials") [2]. The frequency modulation of vibrato is described in [2, 3] by four major characteristics: rate, extent, regularity, and waveform. Rate and extent correspond to the modulation waveform's frequency and amplitude, respectively. Regularity describes the uniformity of frequency modulation with time, and it increases with level of training for opera singers. Mathematically, we model vibrato frequency modulation as a sinusoidal waveform [3].

Though it is a not a well understood phenomonon, vibrato is known to result from a combination of physiologic factors during singing, including the opening and closing of the pharynx and vocal cords, sub-glottic breath pressure, and the actions of the singer's muscles [1, 4, 5]. In western opera singing, vibrato production appears to be mainly due to pulsating contractions of the cricothyroid muscle. Vibrato that is often heard in popular singing and in singing by oriental cultures seems to be caused by pulsation of the subglottal pressure [5].

Vibrato is generally considered to be the result of unconscious effort by well trained singers [4]. A recent study confirmed the important role of the auditory feedback control loop—termed pitch-shift reflex—in vibrato production and regulation [6]. This is in opposition to consciously modulating one's pitch, as in a vocal trilling or by pulsation of the diaphragm. Figure 1 illustrates the variation in fundamental frequency about the mean with straight tone, vibrato tone, and trilling in singing.

Vibrato rate varies greatly among singers due to a number of factors, including level of

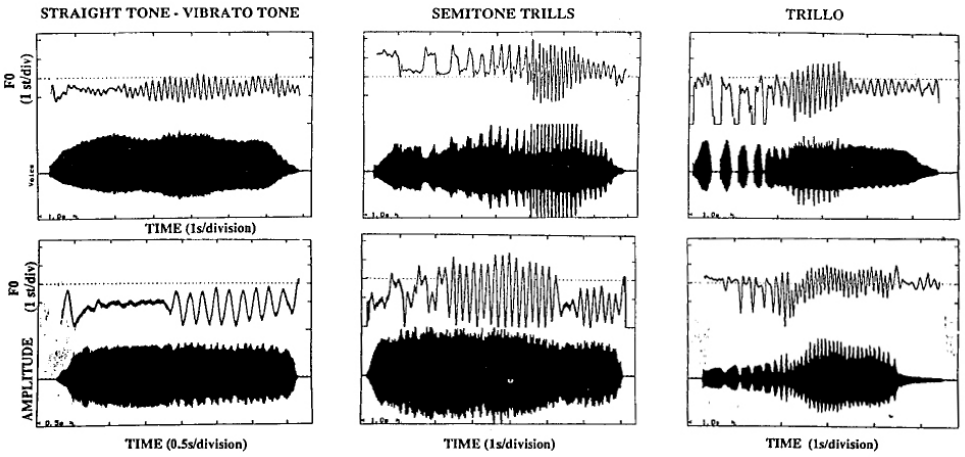


Figure 1: Examples of fundamental frequency F_0 and amplitude variation during straight tone, vibrato, trill, and trillo singing [5]

training, age, personal health, musical style, and cultural preferences [5]. As described in [1], however, it can be broadly categorized as being correctly and incorrectly executed. The following are characteristics of the latter group:

- Wobble — a wide and slow vibrato; usually a trait of older singers that is caused by lack of resistance to breath pressure.
- Overly fast — a tremor usually caused by tongue or body stiffness, improper adduction of the cords, or lack of breath support.
- Straight tone — a lack of vibrato; usually due to excessive and intentional pressure at the glottis.

Sundberg's treatise on singing [7] states that the rate of frequency modulation is constant for a given singer. However, a comprehensive study [2] revealed that the rate tends to increase by about 13% towards the end of a tone. Also, the inter-tone rate variation varied by $\pm 10\%$ about the mean rate for a given singer. The mean vibrato rate among the professional singers tested in the study was 6.1 Hz, with an inter-singer variation of $\pm 10\%$. In the realm of opera singing, it is generally accepted that vibrato has a rate roughly within this range [5]. And for a trained singer, the rate should remain fairly constant—as per Sundberg's statement—regardless of pitch and volume.

Vibrato extent (modulation amplitude) is not so easily characterized, as it is a function of the particular vocal harmonic (or formant) under consideration. Analyzing only the fundamental harmonic of 10 professional singers singing at 25 pitches each, it was found that individual vibrato extent varied between ± 34 cent and ± 123 cent (where ± 100 cent corresponds to one semitone of variation) [8]. A range of ± 71 cent was found about the mean across all

singers and pitches. Extent tends to increase with vocal loudness. One study showed the extent of one singer's vibrato to increase from ±60 cent in pianissimo singing to ±100 cent at fortissimo. We note that there are very complex interactions between the factors of rate, extent, and regularity.

1.2 Vocal Formants

Figure 2 below shows the time-frequency spectral distribution for a tenor singing an ascending A-flat major scale on an "ah" vowel. The fundamental frequency of the first note is approximately 208 Hz; that of the last note is approximately 415 Hz. The vibrato is clearly visible as periodic oscillations in frequency of the pitch.

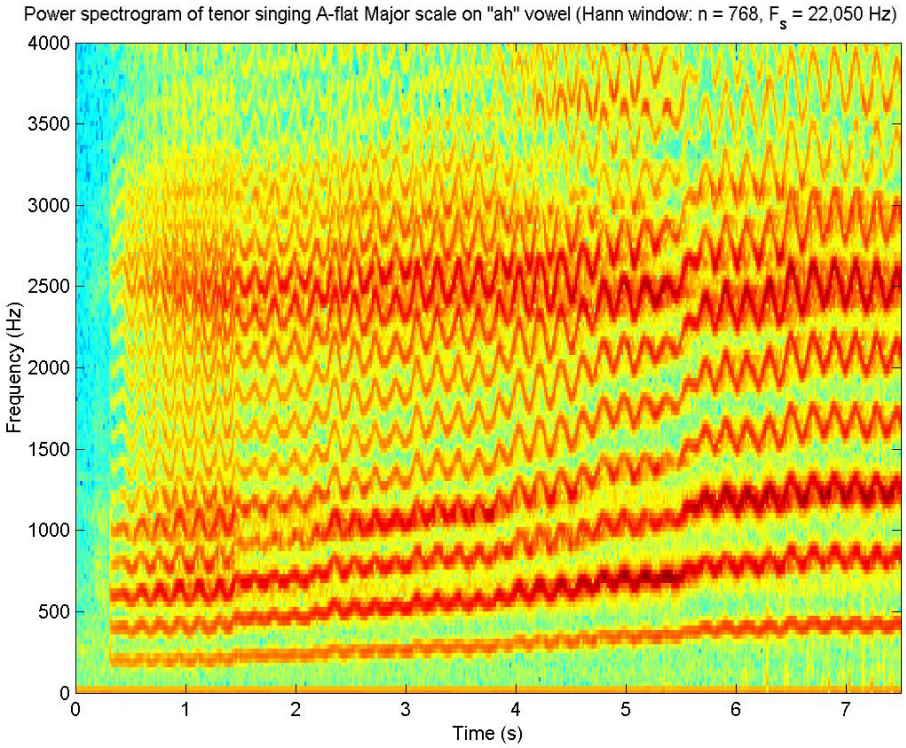


Figure 2: Spectrogram of a tenor singing an ascending A-flat major scale on an "ah" vowel

Formants are the relatively large peaks in the frequency spectrum of a voice that result from resonance within the vocal tract, which is often abstracted as a sound filter [5, 9]. The frequency spectrum of the vocal tract depends on these formant frequencies. Partial frequencies near the formants are stronger than other partials. Since formant frequencies depend on vocal tract shape, it is implied that they will be modulated if vibrato also involves modifications of the vocal tract.

In singing, as in speech, different formants characterize different vowel sounds. A trained

operatic voice contains more higher frequency formants than an ordinary speaking voice. Most opera singing has a distinctive strong formant around 3 kHz, often referred to as the *singer's* formant or the third formant (because it is usually the third major spectrum peak), which is not present in ordinary speech [3]. A fourth formant (above the frequency of the third) is occasionally also present. Figure 3 compares the frequency distribution of an operatic tenor voice with that of ordinary speech and an orchestra. The spectrogram in Fig. 4 shows how the formants of an operatic baritone vary with vowel sound.

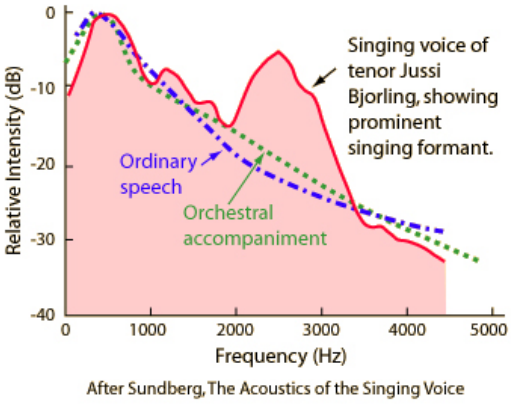


Figure 3: Comparison of operatic tenor's vocal spectrum with speech and orchestra [7]

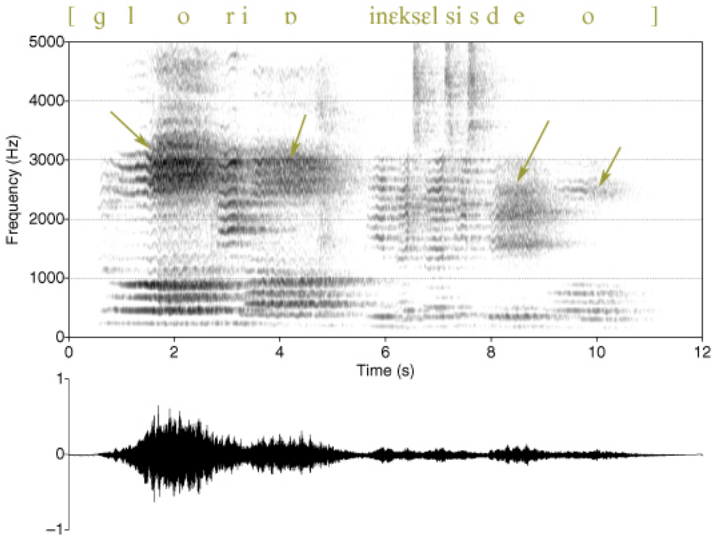


Figure 4: Spectrogram showing formant variation with vowel sound [9]

A formal discussion on formants would be too lengthy to discuss in this space. It suffices to say that for any given voiced pitch with some fundamental frequency F_0, its harmonics at **integer multiples** $n \cdot F_0$ will also be present to some degree. The most significant peaks

among these are called the formants. As noted earlier, vibrato amplitude varies with the harmonic being analyzed in a voice. This is evident from Fig. 2. However, as an approximate rule, the amplitude of vibrato centred at frequency f is a fixed percentage of f.

2 Method

Vibrato is marked by a variation of frequency with time in a one-dimensional signal. Hence, it is logical to perform vibrato detection in the time-frequency domain of the non-stationary signal. And since the magnitude of a time-frequency domain representation (called the spectrogram) of a signal can be treated as an image, image processing and analysis techniques are at our disposal in this domain. In particular, we note that vibrato appears as approximately sinusoidally varying high power bands in the spectrogram.

The purpose of our program is to determine the vibrato rate, the vibrato extent, and the significant formants in a singer's voice. We note that the vibrato extent is a function of formant frequency. Our program incorporates four major components that operate in succession:

1. Short-Time Fourier Transform — transform the sampled voice signal into a time-frequency (spectrogram) representation;
2. Contrast enhancement and edge detection — enhance the appearance of the vibrato in the spectrogram;
3. Hough-Radon Transform — detect sinusoids in the spectrogram;
4. Thresholding and clustering — detect and categorize features in the Hough-Radon domain.

We proceed to investigate each of these steps at a lower level of abstraction.

2.1 Time-Frequency Representation: Short-Time Fourier Transform

There are many transforms that take a 1-D signal $x(t)$ to a 2-D time-frequency distribution representation $TFD_x(t, \omega)$. Several important critera that these transforms should satisfy are given in [10]. Comprehensive reviews of time-frequency distributions and analysis are presented in [11, 12].

We use the Short-Time Fourier Transform (STFT) as the basis for our time-frequency analysis. The STFT is generated by applying the Fourier Transform to a sequence of windowed

signal segments. For a time signal $x(t)$ it is defined as

$$STFT(t,\omega) = \int_{-\infty}^{\infty} x(\tau)w(\tau - t)e^{-j2\pi ft}d\tau, \tag{1}$$

where ω denotes frequency and $w(t)$ is the chosen windowing function, which we choose to be a Hanning function. The purpose of the windowing function is to break $x(t)$ into quasi-stationary segments. Computationally, a Hanning window $w(k)$ of width K samples is defined as

$$w[k] = \frac{1}{2} - \frac{1}{2} \cdot \cos\left(\frac{2\pi k}{K-1}\right), \quad k = 0, \ldots, K-1. \tag{2}$$

Following [13], the STFT is implemented discretely by computing the Fast Fourier Transform of overlapping windowed signal segments. The k^{th} windowed segment of the discretized signal $x(n)$, $n = 0, \ldots, N-1$, is

$$x_k(n) = x(n) \cdot w(n - kP), \quad k = 0, \ldots, M-1, \tag{3}$$

where P is the number of samples to overlap adjacent windows. The number of overlapping segments to be Fourier transformed is $M = \lfloor (N-P)/(K-P) \rfloor$. In our implementation, we choose the window length $K = N/2$ in order to maintain a certain degree of continuity in the STFT. In this case, the overlapped Hanning windows possess the property of summing to unity (they form a *partition of unity*). Though not important for our purposes, this property leads to the invertibility of the STFT.

It is crucial to correctly choose the window length K [13]. The length must be small enough to ensure that the windowed signal is relatively stationary, but long enough to provide sufficient frequency bandwidth for analysis. A short (long) window provides good (poor) time localization, but poor (good) frequency resolution. Resolution in the time-frequency plane is limited by the uncertainty principle [13]:

$$\Delta t \cdot \Delta\omega \geq 1/2, \tag{4}$$

where Δt and $\Delta\omega$ are the time extent (duration) and frequency extent (bandwidth) of $x(t)$ and its Fourier transform, respectively. Both time and frequency resolution cannot be made arbitrarily high simultaneously. In practice, we find that window lengths of between 512 and 1024 samples are optimal for our test signals, which are sampled at 22,050 Hz.

Figure 5 shows the spectrograms of the tenor's A-flat major ascending scale using four different window widths. Note that tradeoffs in both time and frequency resolutions as window width varies.

2.2 Time-Frequency Vibrato Enhancement

Let us clarify the rather vague statement of this step of the program. It is apparent from Fig. 2 that each sinusoidal vibrato signal is spread in the time-frequency plane. The sinusoids

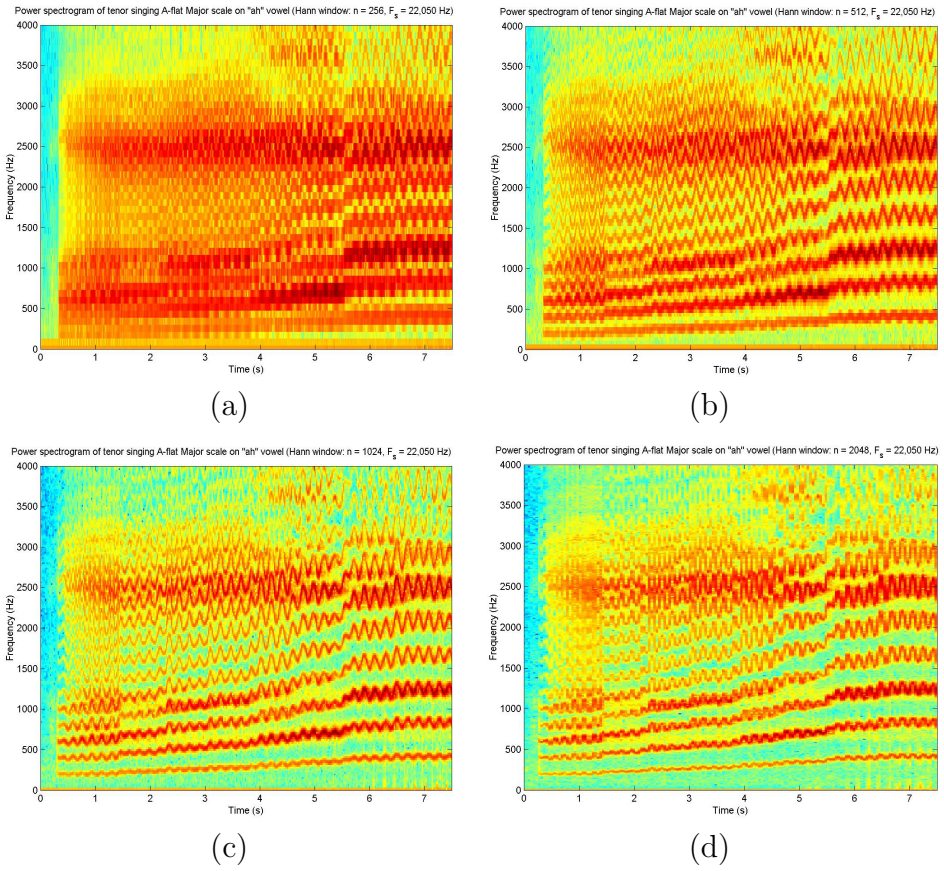


Figure 5: Spectrogram of tenor scale using Hanning windows of size (a) 256, (b) 512, (c) 1024, (d) 2048

have finite area with associated thicknesses: they are not ideal thin curves. This is a result of both the inherent resolution limitation in the time-frequency plane (Eq. 4) and the true spread of vocal frequencies.

We would like to identify each of these thick lines as a single vibrato component, as opposed to several. Reassignment methods, which remap the time-frequency plane in order to localize spread-out structures, could help to localize the vibrato spreading [14, 15, 16]. We have chosen not to implement time-frequency reassignment, though we show the effect of applying one such algorithm on an example signal below. Figure 6 shows an perfectly sinusoidally modulated signal, along with its spectrogram and its Fourier Transform. Figure 7 shows the reassigned spectrogram of the same signal. These signals and plots were generated using the Time-Frequency Toolbox[1].

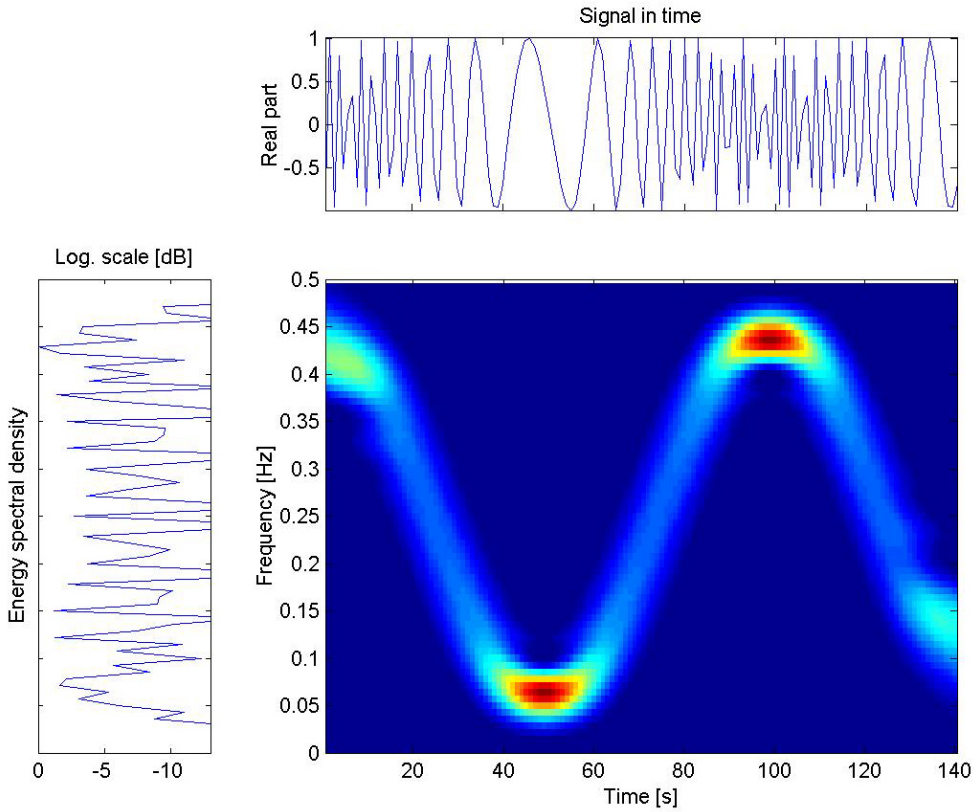


Figure 6: Spectrogram of sinusoidally modulated time signal

[1]Time-Frequency Toolbox (for use with MATLAB) by F. Auger, P. Flandrin, P. Gonçalvès, and O. Lemoine (1995-96) available at http://tftb.nongnu.org/

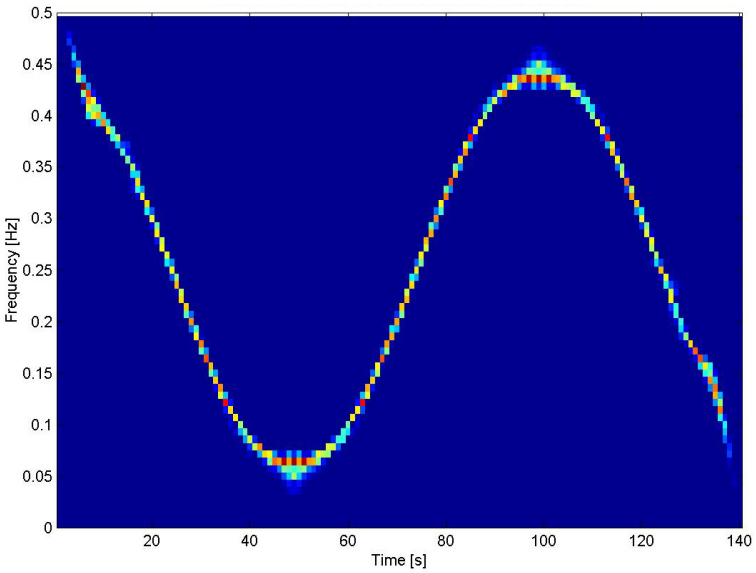


Figure 7: Reassigned spectrogram of sinusoidally modulated time signal

2.2.1 Contrast Enhancement

High intensity points appear brighter than low intensity points in the spectrogram. We increase spectrogram contrast in order to enhance the contours of vibrato structures. This achieved by stretching the image histogram to the full available range of gray-scale values. This simple step is found to aid in accentuating the vibrato from the spectrogram's background information. We also experimented with gamma correction filters, but these only further obscured the vibrato.

2.2.2 Canny Edge Detection

We apply Canny's method for edge detection to the spectrogram images in order to discriminate the boundaries of the vibrato signals. Canny's algorithm uses multidirectional derivatives, multiscale analysis, and optimization procedures in order to detect edges [17]. The method is characterized by low probabilities of false edge detection and of erroneously detecting edges, good localization of detected edges, and production of a single output from a given edge.

The purpose of edge detection is to improve localization of features in the Hough-Radon space, as will be discussed. We have found that applying the Hough-Radon Transform to edge-detected versions of spectrograms results in improved sinusoidal parameter detection. This assumes that the detected edge correspond spatially with the sinusoids of the spectrogram.

2.3 Vibrato Detection: Hough-Radon Transform

The Hough-Radon transform (HRT) is described in [18], where it is applied to the detection of linearly frequency-modulated signals in the time-frequency domain representations of knee-joint vibration sounds. The Hough transform itself, which is equivalent to the Hough-Radon transform when applied to binarized images, is described in detail in [19]. The Radon transform is a less general version of the HRT, and is only used to detect straight lines.

Let us assume that each vibrato signal has infinite time duration. Let us then parameterize a given vibrato signal as a sinusoid in the time-frequency plane:

$$\omega_{(a,b,c,d)}(t) = a \cdot \sin(b \cdot t + c) + d, \tag{5}$$

where we define the following parameters of the vibrato:

- a — extent, or amplitude,
- b — rate, or osciallation frequency,
- c — initial phase offset,
- d — harmonic/baseline frequency.

The HRT algorithm stores a 4-D floating-point data structure B, with each dimension corresponding to one of the parameters a, b, c, d. We refer to element $B(a_i, b_i, c_i, d_i)$ as the *bin* for quadruple (a_i, b_i, c_i, d_i). After applying the transform, each bin $B(a_i, b_i, c_i, d_i)$ stores the integral of the spectrogram $S(t, \omega)$ along the $t - \omega$ planar curve $\omega_{(a,b,c,d)}(t)$:

$$B(a_i, b_i, c_i, d_i) = \int_{t=0}^{t=t_{\max}} S(t, \omega_{(a,b,c,d)}(t))\, dt \tag{6}$$

Thus, the value $B(a_i, b_i, c_i, d_i)$ is a measure of the strength of the sinusoid $a_i \cdot \sin(b_i \cdot t + c_i) + d_i$ in the spectrogram. We refer to the (a, b, c, d) parameter space as the Hough-Radon space.

We note that the four Hough-Radon domain parameters, as well as t and ω must be quantized appropriately, so as not to exceed memory and computation time limitations. In addition, intelligent choices must be made for the upper and lower bounds of the parameters. These choices relate to the acoustics of vibrato, as discussed earlier. Quantization is discussed in section 3.1.

2.3.1 Implementation

The HRT is implemented as follows:

1. Bound and quantize the Hough-Radon and time-frequency spaces,
2. Loop through each point (a_i, b_i, c_i, d_i) in the Hough-Radon space:

   For each point t_i in the time domain:

   (a) Compute $\omega_{(a_i,b_i,c_i,d_i)}(t_i)$,
   (b) Increment the bin $B(a_i, b_i, c_i, d_i) = B(a_i, b_i, c_i, d_i) + \omega_{(a_i,b_i,c_i,d_i)}(t_i)$.

Another possible implementation is as follows: For each point (t_i, f_j) in the spectrogram with a non-zero value, let a, b, c equal each of their allowed quantized values. Solve for the value of the remaining parameter via $d = \omega_j - a_i \cdot \sin(b_i \cdot t_i + c_i)$ and round it to the nearest quantized value. Then increment the bin value by $B(a_i, b_i, c_i, d_i) = B(a_i, b_i, c_i, d_i) + \omega_{(a_i,b_i,c_i,d_i)}(t_i)$.

We normalize the spectrogram values to the range $[0, 1]$ prior to performing the HRT. A relatively high intensity bin value $B(a_i, b_i, c_i, d_i)$ indicates the presence of a vibrato signal $\omega_{(a_i,b_i,c_i,d_i)}(t_i)$. As described in [18], some form of thresholding must be applied to the bin values in order to distinguish the presence of vibrato from background information.

2.3.2 Implementation using Template Matching

The implementation of the HRT described in section 2.3.1 requires that the *sin* function be evaluated at every time location for every choice of parameters. We hypothesized that this was a waste of computational effort and that the HRT implementation could be sped up via template matching. Rather than calculating $\sin(b_i \cdot t + c_i)$ at every point, we index a high resolution 1-D sinusoidal template array. In our implementation, the template had 10,000 sample points for one period of a sinusoid. The value $\sin(b_i \cdot t + c_i)$ is retrieved by modular indexing of the template array.

Over several comprehensive tests, computation time of the HRT using the template matching implementation was found to (surprisingly) take on average 8.4% longer than using the original implementation. Thus, the original implementation was used for all testing. This result seems to indicate that the *sin* function in MATLAB is highly optimized.

2.4 Parameter Detection: Hough-Radon Filtering

We have already established that a sinusoidal pattern will appear as relatively high intensity features in the Hough-Radon (H-R) domain. This assumes two things: First, that the true parameters of the sinusoid in the time-frequency plane are within the boundaries of the H-R parameter space. Second, that the H-R parameter space quantization is sufficiently dense to detect the sinusoid.

Detecting high intensity features that represent sinusoids is accomplished by filtering in the H-R domain. The domain is initially thresholded in order to retain high intensity candidate sinusoid parameters. Now suppose that the given parameter vector (a_i, b_i, c_i, d_i) is "close" to (a_i', b_i', c_i', d_i'). (In the work that follows, we quantify "closeness" in the 4-D Euclidean distance sense.) If the nearby H-R bins $B(a_i, b_i, c_i, d_i)$ and $B(a_i', b_i', c_i', d_i')$ both surpass the threshold, then it is possible that their associated parameters should be clustered together in order to designate the parameters of a single sinusoid. In other words, both bins may indicate the presence of sinusoids from the same vibrato signal. Thus, we follow H-R domain thresholding with clustering.

2.4.1 Thresholding

Thresholding is used to distinguish high intensity sinusoidal parameters in the H-R domain. The threshold is computed based on the histogram $P_{\mathrm{HR}}(l)$ of the 4-D H-R domain. Let μ_{HR} and σ_{HR} denote the mean and standard deviation of the histogram. For many of our test cases, thresholds of $T = \mu_{\mathrm{HR}} + m \cdot \sigma_{\mathrm{HR}}$, where m ranges from 5 to 10, are good at discriminating sinusoids in the H-R domain. We note that the H-R histogram means μ_{HR} tend to be quite low for our test cases, as the majority of points in the H-R domain do not correspond to sinusoids and thus have very low intensities. Several examples of these histograms are given in section 4.

Another way that we choose the H-R threshold is by searching for the gray-value T below which a certain proportion of the histogram's weight is distributed. Thus, we may choose the threshold as the minimum value of T that satisfies $\sum_{l=0}^{T} P_{\mathrm{HR}}(l) \geq p$, where P_{HR} is the normalized histogram and p is the cutoff probability. We choose p to be appropriate to distinguish vibrato. In our test cases, p ranges from 0.95 to 0.995 and it depends on nature of the H-R domain.

2.4.2 K-Means Clustering

We apply the iterative K-means clustering algorithm [19] to the H-R domain subsequent to thresholding. This step is used to cluster together all close H-R points (i.e. bins) that denote the presence of a single vibrato sinusoid parameter vector. The algorithm terminates with the centres of K cluster domains. These centres are minimize the sum of the squared distances between all points of a cluster and the cluster's centre. The value K is set to equal the number of suspected sinusoids in the H-R domain. In practice, this number is determined by visual inspection of the time-frequency representation.

3 Sinusoid Detection

Here we elaborate on how the methods of section 2 are used to detect vibrato.

3.1 Parameter Quantization

Given that it is a 4-D data structure, the full H-R domain grows very quickly in size as the search space parameter resolution is increased. More significantly, computation time becomes a limiting factor. Timings of the HRT algorithm versus the number of quantization bins used are given in Table 1. As expected, there is a linear dependence between number of bins and the computation time.

Number of H-R bins	**Average run time (s)**
1	0.015
11	0.078
21	0.125
41	0.235
81	0.437
161	0.844
321	1.609
641	3.235
1281	6.500
2561	13.079
5121	25.641

Table 1: Average HRT algorithm timings (for four runs at each bin level) on a 3 GHz Pentium 4 with 1 GB of RAM.

The majority of our test cases consisted of sound files with duration of about 2 to 5 seconds sampled at 22,050 Hz. Let N_t and N_f denote the number of pixels along the time and frequency axes of the resulting spectrogram, respectively. As previously stated, $N_t = \lfloor (N - P)/(K - P) \rfloor$, where N is the number of sound samples, K is the window length (usually either 768 or 1024 samples), and $P = K/2$ is the window overlap. The number of frequency samples is $N_f = K/2$. Thus, our spectrogram images were no in general no bigger than roughly $N_t \times N_f = 300 \times 500$ pixels. The majority of our spectrograms have frequency axes limits of 0 to either 4,000 Hz or 8,000 Hz.

Quantization for each parameter is described separately below. We only give approximate ranges for the parameter search space resolutions, as the actual values used depended highly on the examples investigated. When the resolution of one parameter is chosen to be particularly high, there is often a trade-off in the resolution of another parameter. This discussion only applies to the quantization used for actual (recorded) voice data, as opposed to synthetically generated data.

- a — **extent**: Amplitudes of sinusoids are observed vary between 10 Hz and 110 Hz in the time-frequency representations of our examples. Also, extent generally increases with the harmonic frequency (d) of the sinusoids. The relationship between a and d is complex, and analysis of these two parameters on a real example is given in section 4.2. If we are analyzing vibrato at a known partial frequency d, then we quantize a to up to 100 values in the range $[a_{\min}, a_{\max}]$, where $a_{\min}$ and $a_{\max}$ are intelligently chosen to encompass the expected range of vibrato extent at partial d. Otherwise, we allow a to vary between the full 10 Hz to 110 Hz range.

- b — **rate**: The vibrato rates for our examples (all trained opera singers) are known to be safely within the range $b_{\min} = 5.5$ Hz and to $b_{\max} = 8.5$ Hz. This range is evenly quantized into between 10 and 50 evenly spaced values for the majority of our examples.

- c — **initial phase**: This can be anywhere in the range 0 to 2π. However, we note that all sinusoids in the spectrogram of a single voice are in phase. Thus, parameter c is constant for each test. Our minimum/maximum range for c is estimated based on visual inspection of the spectrogram. Generally, this range is kept quite small and it is divided into no more than 10 values.

- d — **harmonic frequency**: It is known that partials appear at all integer multiples of the sung note's fundamental frequency. The fundamental frequencies in our examples ranged from between about 200 Hz to 1000 Hz. Given that the maximum search space for d is usually $[0, 4000$ Hz$]$, the number of partials at which sinusoids need to be detected may range from 4 to 20. When the note's fundamental frequency is known, we choose the search space for d to include the frequencies of the suspected partials. When no information is given, we search the entire frequency range using between 10 and 100 values for d.

3.2 Parameter Space Optimization

In the discussion that follows, we treat the problem of determining sinusoid parameters in the HRT as an optimization problem. In this context, "searching" the parameter space along certain variables refers to performing HRT integration along the corresponding sinusoids.

Intelligent reduction of the parameter search space can be achieved by considering the nature of vibrato in the time-frequency representations. In general, we begin by coarsely stepping through the parameter space. Once candidate parameters for the vibrato have been found, we increase parameter resolution in the search space around the candidate parameters. This saves the computational effort of finely searching through locations in the parameter space where no vibrato is present. High resolution searches are generally performed on the Canny edge filtered spectrograms, whereas low resolution searches are performed on the unfiltered spectrograms.

The vibrato sinusoids of one voice are known to be in phase (constant c) and to all possess an approximately constant rate (b). Thus, we search for single values of both b and c that maximize the HRT. The vibrato sinusoids can have varying (although correlated) values of a and d, which we wish to determine. The steps of our strategy are as follows:

1. **1-D search for approximate maximizers in** d — Rather than performing a brute force HRT using every allowable parameter value, we begin by performing the HRT along only the d parameter space. This is done by setting $a = 0$ in order to disregard all parameters except for d in the HRT. In other words, we perform integration along horizontal lines. Peaks in the resulting 1-D HRT then indicate the approximate locations of the partials.

2. **4-D search for approximate maximizer in** c — The search space in the HRT is broadened to include all variables a, b, c, d, with searching along d performed using the candidate positions of step 1. The resolutions along a, b, c are low in this step. We determine a single value for the initial sinusoid phase shift by searching for a global maxima in the H-R domain along only the c direction. The search is performed by transforming H-R domain into a 1-D function by summation along parameters a, b, c.

3. **4-D search for true maximizer in** c — We repeat the search of step 2 starting around the candidate phase shift value found. The resolution along c is very high. We repeat the 1-D maxima search to find the exact value for c that maximizes the HRT. **This is the final value determined for the vibrato phase shift.**

4. **3-D search for approximate maximizer in** b — The search of step 2 is repeated using the single value for c determined in step 3. We increase the resolution of the search space in parameter d. More accurate estimates of the partial frequencies are found by searching for maxima along d. Also, approximate values are determined for a, b.

5. **3-D search for true maximizer in** b — The search of step 4 is repeated using increased resolution in b and the more finely tuned approximations for maxima along d. We find the exact value of b that maximizes the HRT. **This is the final value determined for the vibrato rate.**

6. **2-D search for true maximizers in** a, d — The resolution along directions a, d is greatly increased, with parameter values for a chosen to roughly increase with increasing d. This corresponds to our observation that vibrato extent (a) is an increasing function of harmonic frequency (d). **This final high resolution search yields the parameters** a, d **of the vibrato sinusoids present in the spectrogram.**

3.3 Multiresolution Strategy

A multiresolution strategy was devised in order to facilitate the parameter space searching. The HRT is applied to spectrograms of progressively finer resolution: $\{S^1, S^2, \ldots, S^{n-1}, S^n\}$. Here S^n is the original image (highest resolution), S^{n-1} is the image at the next lower resolution, ..., and S^1 is the lowest resolution image. We use $n = 4$ for our experiments and choose to downsample S^i by factor of 2 along each dimension (using bicubic interpolation) to obtain S^{i-1}. We also apply progressively increased Gaussian blurring to the downsampled images.

The underlying assumption behind this strategy is that the approximate parameters of sinusoids are easier to obtain (due to fewer local optima in the search space) at low resolutions. This is because the sinusoids (along which HRT integration is performed) are "thicker" with respect to structures in lower resolution spectrograms. Fine tuning of the parameters is achieved at higher spatial resolutions of the spectrograms.

4 Results and Discussion

We demonstrate the application of our methods to both simulatd and actual vibrato data.

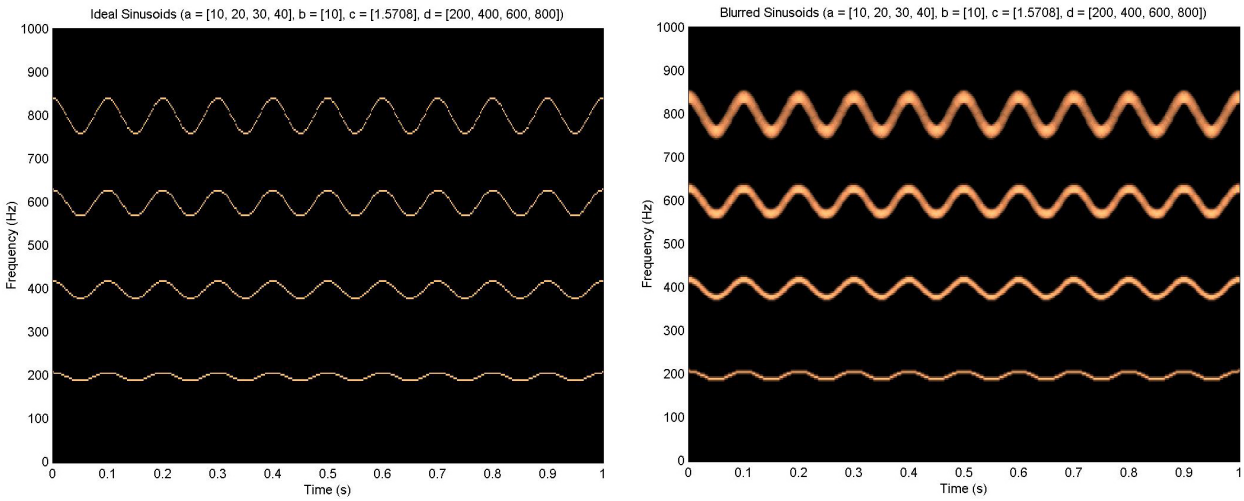
4.1 Simulated Data

The first two examples, shown in Fig. 8, are of simulated spectrograms for which the vibrato parameters b and c are held constant at 10 Hz and $\pi/2$ radians, respectively. The examples have time and frequency scales of $[0, 1]$ seconds (500 samples) and $[0, 1000]$ Hz (250 samples), respectively. Circular averaging is performed on the sinusoids in Fig. 8(b) using circular average filters with radii of 1, 2, 3, and 4 samples.

4.1.1 1-D Hough-Radon Transforms

Figure 9 shows H-R transforms of the ideal sinusoid at 600 Hz in Fig. 8(a) obtained by searching along each parameter a, b, c, d individually. The parameters are varied about the exact sinusoid parameters. Each HRT is computed using 200 parameter values (bins). Observe that the HRT exhibits a global maximum at the precise sinusoid parameter values. The only non-symmetric HRT is that Fig. 9(a) obtained by varying a. For $a \in (26, 30)$ the HRT is convex, while for $a > 30$ the HRT is concave.

Figure 10 shows H-R transforms of the sinusoids in Fig. 8 (a) and (b) obtained by varying parameter d through the entire search space, with a, b, c fixed at 20 Hz, 10 Hz, and $\pi/2$, respectively. The resolution of d is 1 Hz. Observe that the since $a = 20$ for these transforms,



(a) (b)

Figure 8: Simulated vibrato: (a) ideal and (b) blurred sinusoids with parameters $(a, b, c, d) = (10, 10, \pi/2, 200), (20, 10, \pi/2, 400), (30, 10, \pi/2, 600), (40, 10, \pi/2, 800)$

the largest peak corresponds to the sinusoid at 400 Hz. The peaks in the HRT of the blurred sinusoids are less differentiated in height than in the HRT of the ideal sinusoids. Also, the main peak in Fig. 10(b) is not as high as the main peak in Fig. 10(a). We note that sinusoids in real spectrograms are blurred, complicating the task of detection.

Our next example is of the 1-D HRT applied to the simulated data in Fig. 11(a). The sinusoids in this data are identical except in frequency, which range from 6.0 to 10.5 Hz in increments of 1.5 Hz. The 1-D HRT in Fig. 11(b) is obtained by varying parameter b between 4.5 and 12 Hz over 200 steps. It is visually very difficult to differentiate the overlapping signals in Fig. 11(a). However, the HRT clearly depicts the four sinusoids.

The final 1-D HRT example in Fig. 12 is somewhat pathological. Nonetheless, it illustrates an extreme case of a both linearly and sinusoidally modulated sinusoid.

4.1.2 2-D Hough-Radon Transforms

We illustrate the HRT applied to the blurred simulated data in Fig. 8(b). We vary two parameters at a time in these examples. Figure 13 shows three 2-D Hough-Radon transforms obtained by varying each parameter through 100 values about the actual parameters of the 600 Hz sinusoid of Fig. 8(b). The intensities on these plots are shown on a $\log_{10}$ scale in order to bring out subtle features.

The histograms of these 2-D HRTs are shown in Fig. 14. We note that all of the content

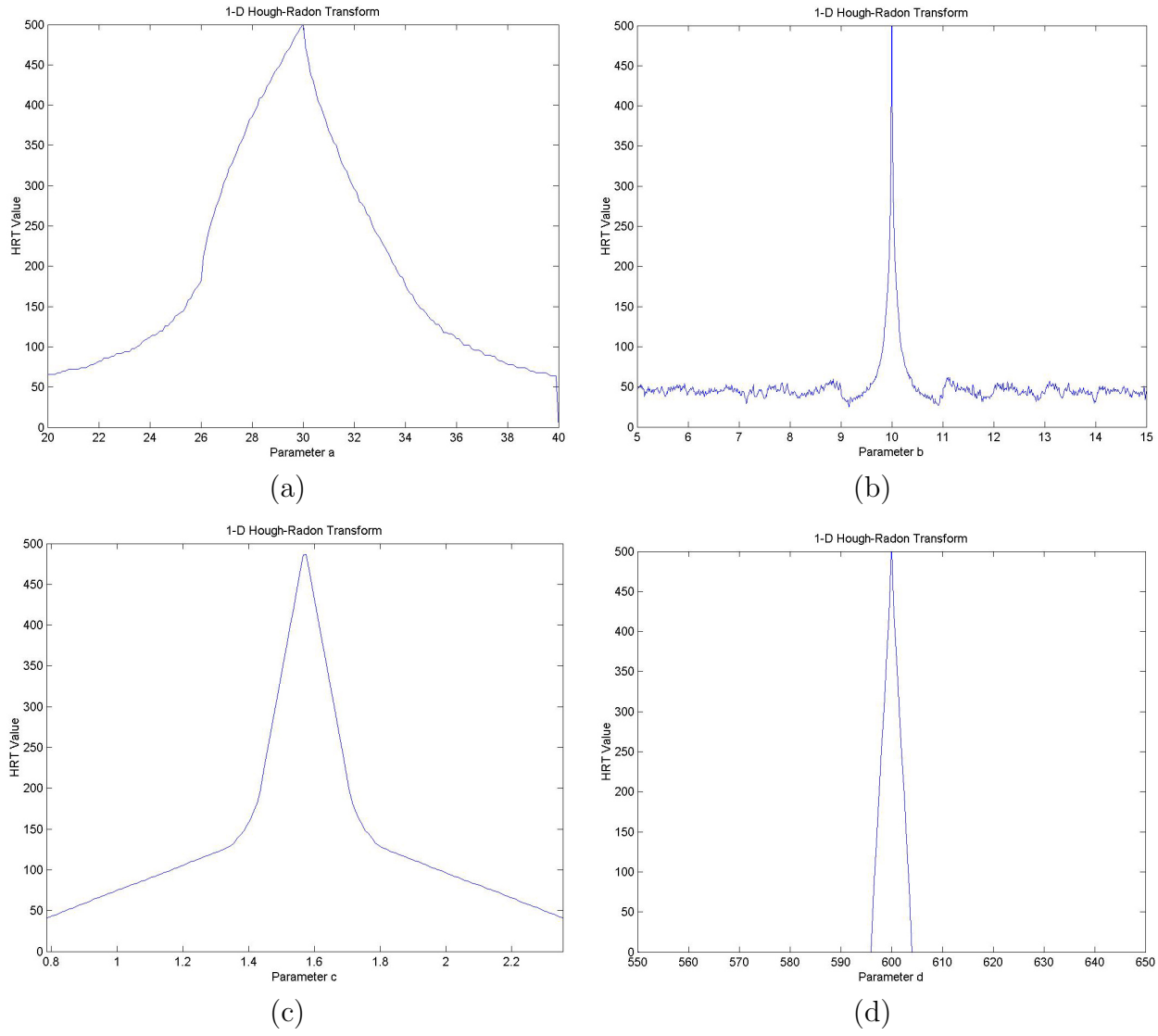
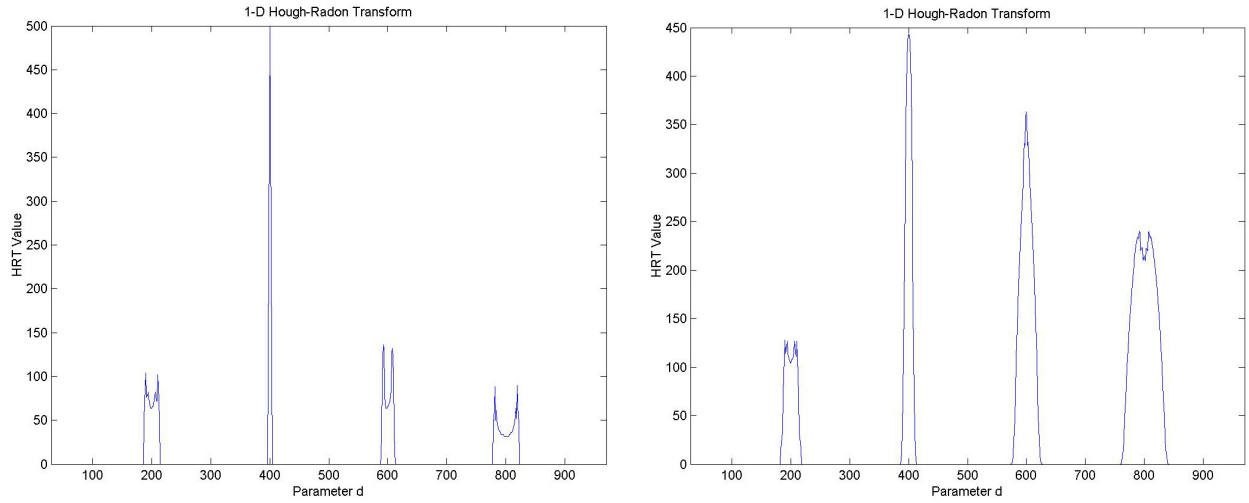
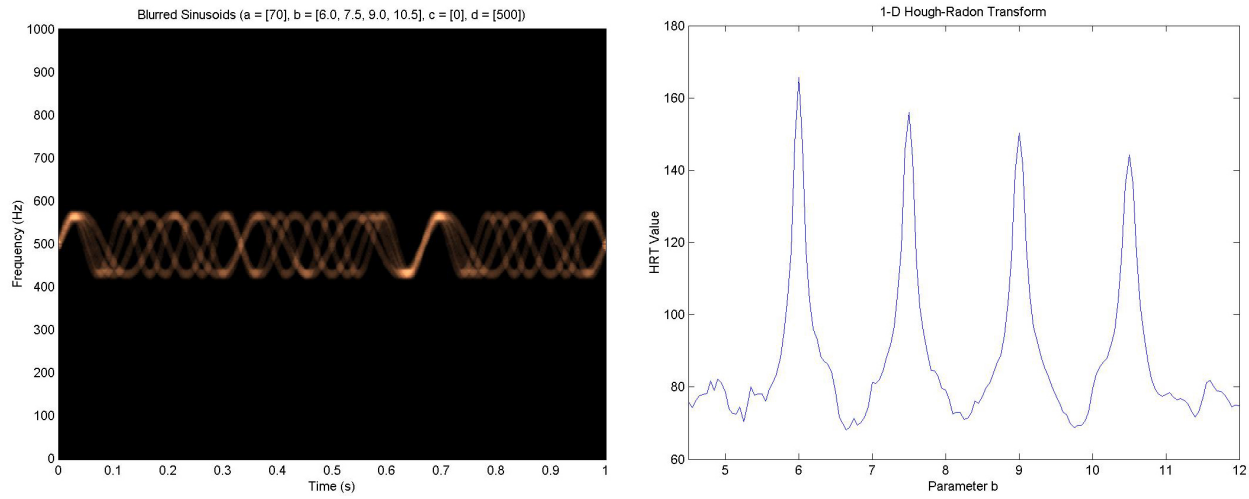


Figure 9: 1-D HRT of simulated ideal vibrato obtained by varying parameters (a) a, (b) b, (c) c, and (d) d individually



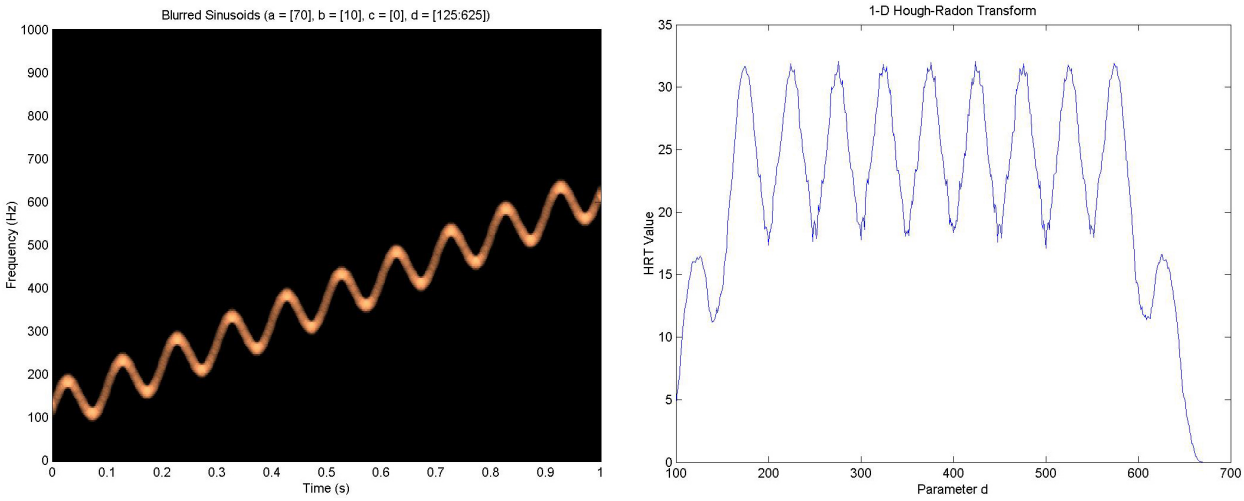
(a) (b)

Figure 10: 1-D HRT of simulated (a) ideal and (b) blurred vibrato obtained by varying parameter d through entire search space



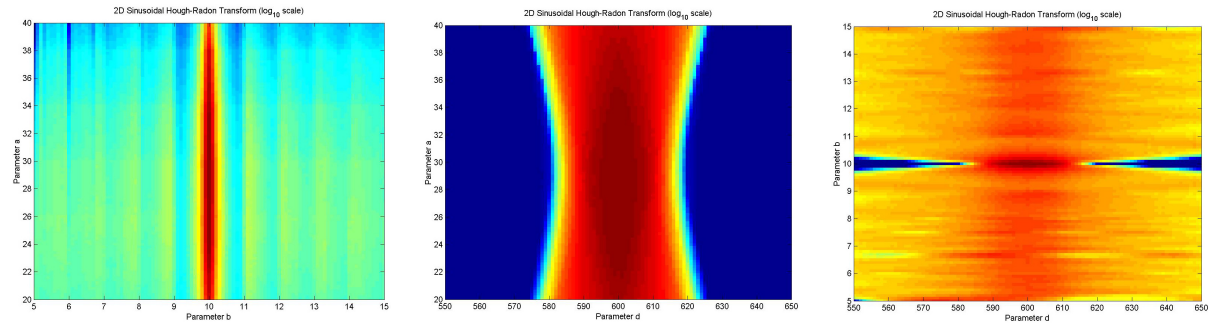
(a) (b)

Figure 11: (a) Simulated (blurred, radius = 3) vibrato with parameters $(a, b, c, d) = (70, 6.0, 0, 500), (70, 7.5, 0, 500), (70, 9.0, 0, 500), (70, 10.5, 0, 500)$; (b) HRT of simulated vibrato by varying b from 4.5 to 12 Hz



(a) (b)

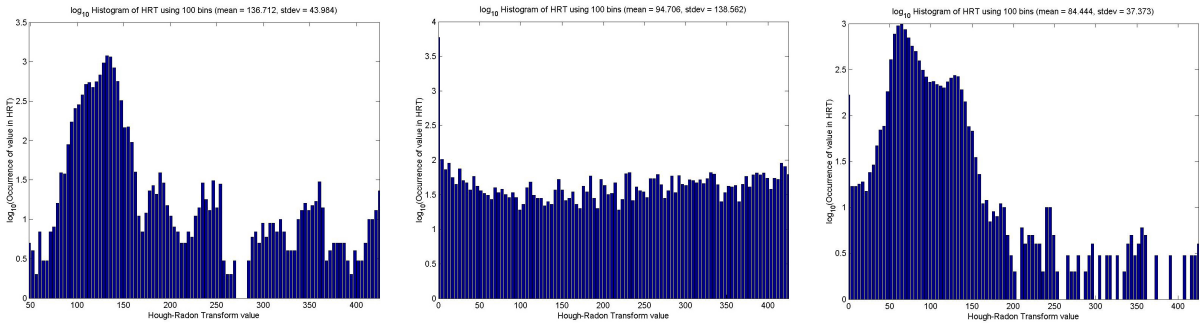
Figure 12: (a) Simulated (blurred, radius = 4) linearly modulated vibrato with parameters $(a, b, c, d) = (70, 10, 0, d)$, where d ranges linearly from 125 to 625 Hz; (b) HRT of simulated linearly modulated vibrato by varying d from 100 to 700 Hz



(a) (b) (c)

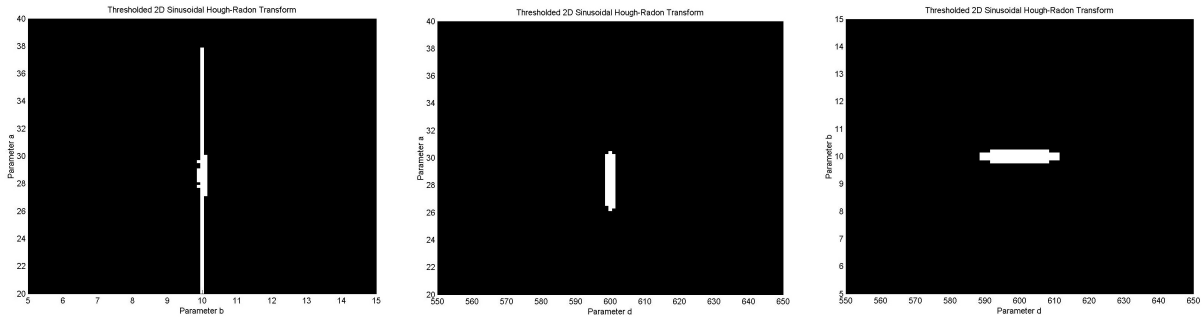
Figure 13: 2-D Hough-Radon transforms on $\log_{10}$ scales obtained by varying parameters through 100 values each, as follows: (a) a, b, (b) a, d, (c) b, d

related to the sinusoids present in the HRTs is at the extreme high intensity end of the histograms. We threshold the HRTs using a probability cutoff $p = 0.99$, as defined in section 2.4.1. The resulting images are shown in Fig. 15.



(a) (b) (c)

Figure 14: Histograms on $\log_{10}$ scales of 2-D HRTs obtained by varying: (a) a, b, (b) a, d, (c) b, d

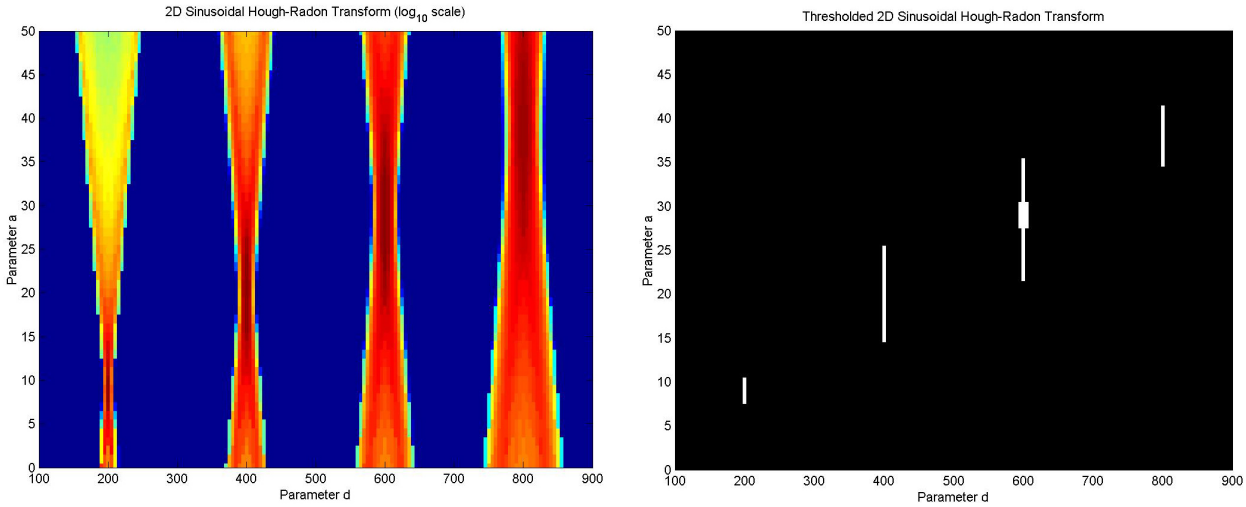


(a) (b) (c)

Figure 15: Thresholded ($p = 0.99$) 2-D Hough-Radon transforms obtained by varying parameters through 100 values each, as follows: (a) a, b, (b) a, d, (c) b, d

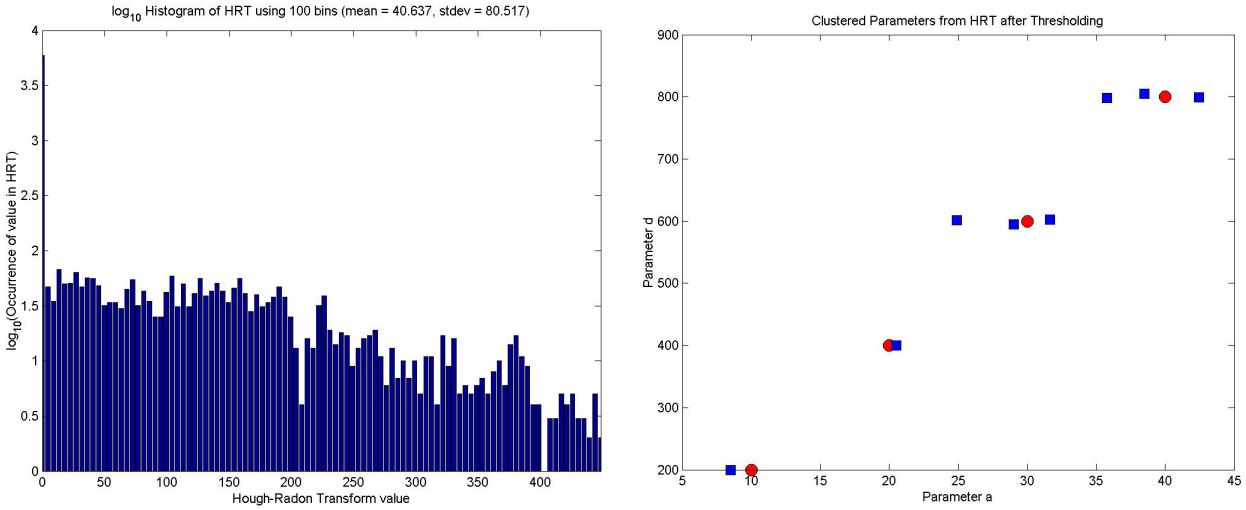
Choosing $K = 1$, we perform K-means clustering on the three thresholded histograms, yielding very good results of $(a = 28.82, b = 10.01), (a = 28.33, d = 600.02), (b = 10.00, d = 600.00)$ for the sinusoid.

For our final 2-D HRT example again deals with the simulated data in Fig. 8(b). We vary the parameters a, d over the range of all permissible values. The computed 2-D HRT has 10,000 points (100 per variable). Figure 16 shows both the $\log_{10}$ HRT and its thresholded version at $p = 0.995$. The threshold at this level is computed to be 385.07. The histogram of this 2-D HRT and the results of K-means clustering (using $K = 8$) are shown in Fig. 17.



(a) (b)

Figure 16: (a) $\log_{10}$ scaled 2-D HRT and (b) thresholded ($p = 0.995$) 2-D HRT obtained by varying parameters a, d through 100 values each



(a) (b)

Figure 17: (a) Histogram of 2-D HRT obtained by varying parameters a, d; (b) Results of 8-means clustering on thresholded 2-D HRT (true parameter values are shown using red circles, experimental results are shown using blue squares)

4.2 Real Data

We illustrate the sinusoidal Hough-Radon for vibrato detection on the spectrogram data in Fig. 18 of an operatic baritone singing an A (fundamental frequency 213 Hz) and a D (fundamental frequency 270 Hz) on an "ah" vowel. We perform segmentation (with the help of Adobe Photoshop 7.0) to obtain comprehensive manual measurements on the first spectrogram (Fig. 19(a)). These measurements are used gold standards against which we compare our program's vibrato detection results. The vibrato rates (b) in both audio files are computed manually to be 7.5 Hz.

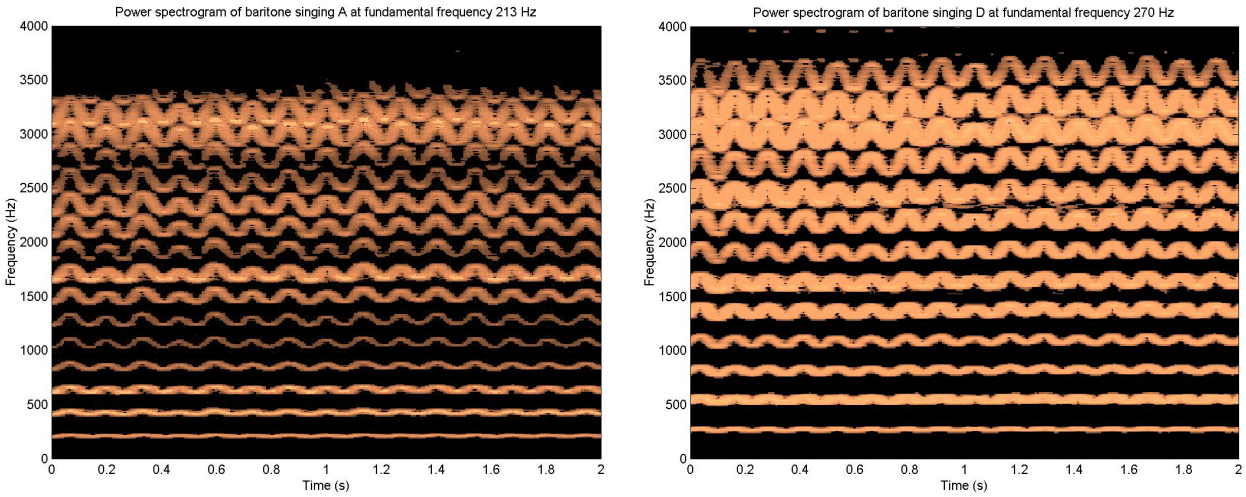
The manual measurements include the following measures of the individual vibrato components at the varying harmonic frequencies:

- mean intensity (0-255 scale),
- total area on time-frequency plane (dimensionless),
- total frequency span (Hz),
- width in frequency (Hz),
- maximum peak-to-peak amplitude (Hz),
- minimum peak-to-peak amplitude (Hz).

The peak-to-peak amplitudes given are equal to twice the extent of the vibrato. Both maximum and minimum values are given for amplitude, as each vibrato signal appear to be composed to both a low and a high amplitude wave. (This indicates that the vibrato signals are in fact not pure sinusoids.) We note that there are clear correlations between all of these variables and the harmonic frequency of the vibrato being analyzed. Vibrato intensities are particularly strong at the four strong partials frequencies, as seen in the spectrogram analysis. Observe that the vibrato amplitude (extent, a) increases nearly linearly with harmonic frequency.

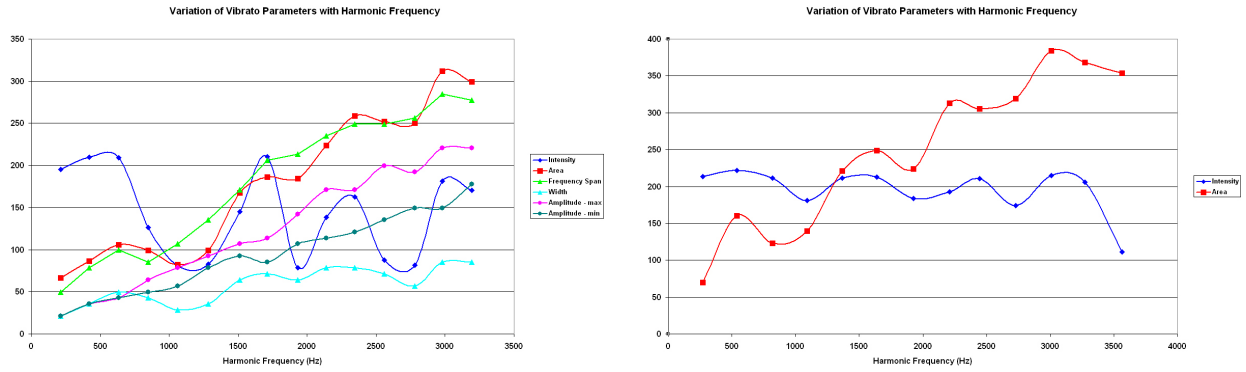
4.2.1 Multiresolution Spectrograms

Figure 20 shows the four spectrogram resolutions used for the analysis of the 213 Hz baritone data. Downsampling (by 2 in both time and frequency axes) was done using bicubic interpolation. A 4×4 Gaussian blur with standard deviation of 4 was applied to the spectrograms at each resolution prior to downsampling.



(a) (b)

Figure 18: Spectrograms of a baritone singing an (a) A (fundamental frequency 213 Hz) and a (b) D (fundamental frequency 270 Hz) on an "ah" vowel



(a) (b)

Figure 19: Gold standard (manual) measurements of spectrograms with baritone singing (a) A (fundamental frequency 213 Hz) and (b) D (fundamental frequency 270 Hz)

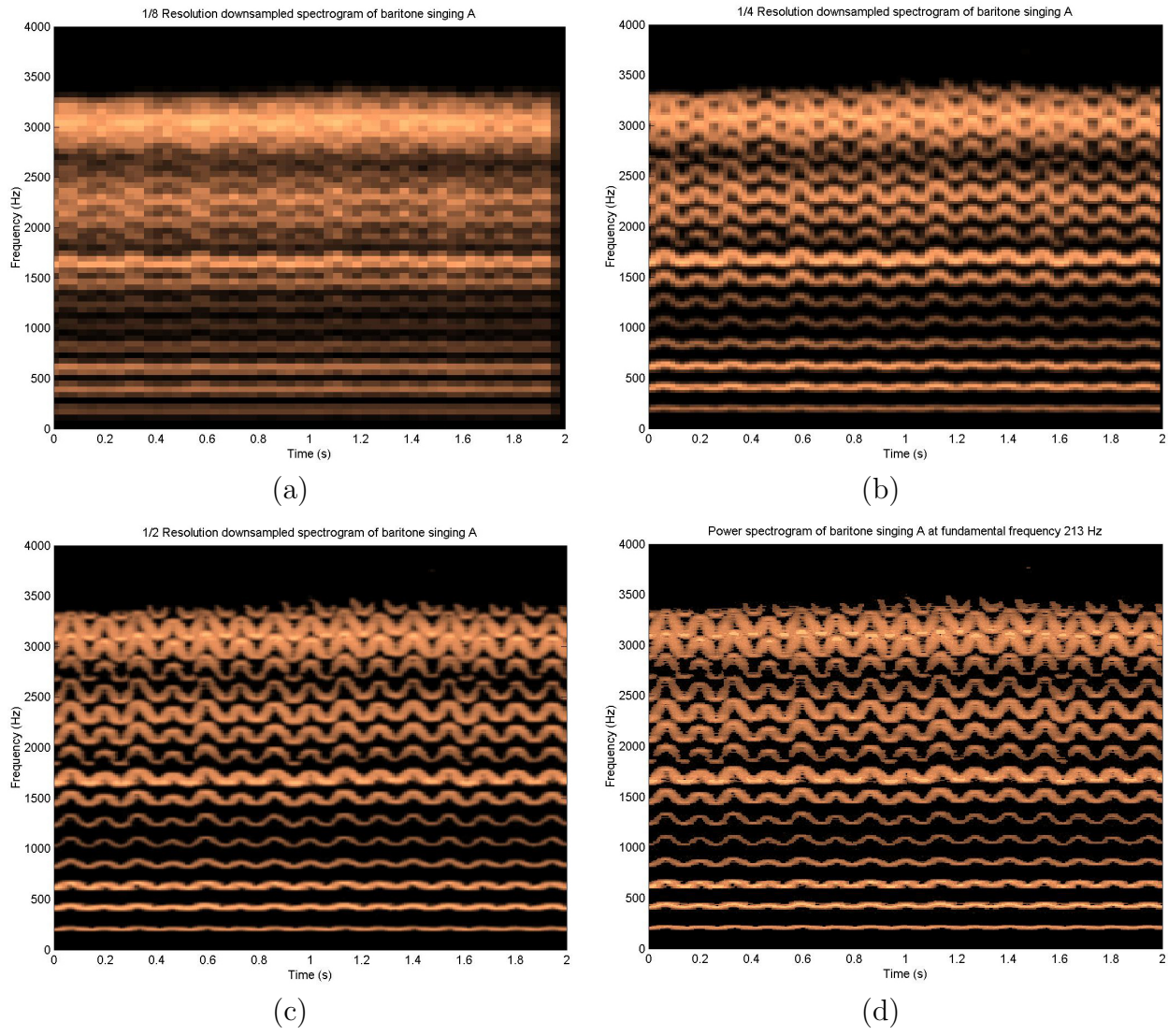
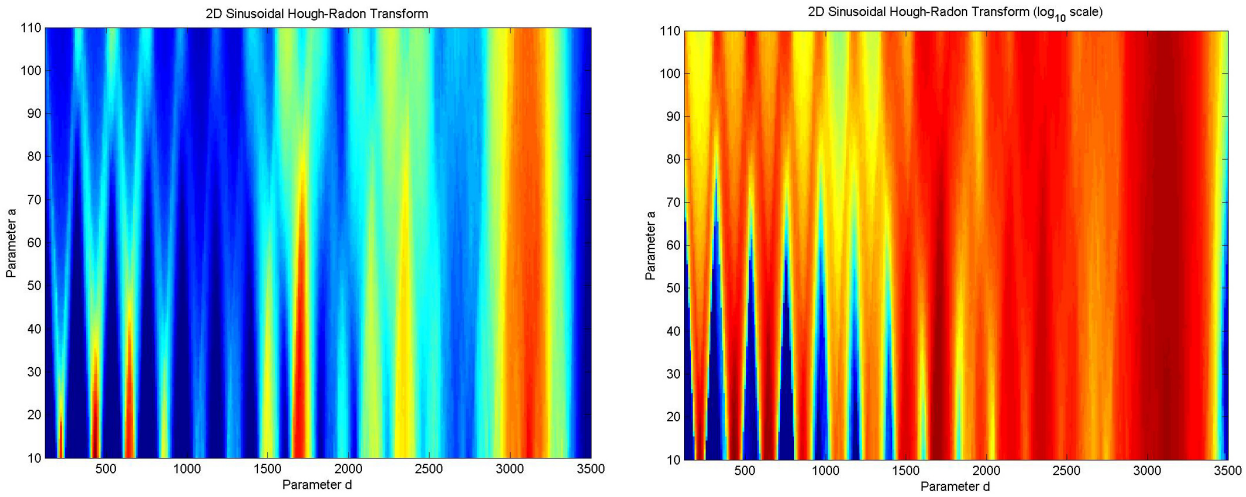


Figure 20: Multiresolution images of baritone singing A. Time-frequency plane resolutions are: (a) 72×55, (b) 142×108, (c) 282×214, (d) 562×427

4.2.2 Hough-Radon Transforms

We employ the strategy described in section 3.2 in order to first obtain the parameters b, c for the vibrato signals. The values of these two parameters are determined to be $b = 7.5$ (as measured by manually) and $c = 3.14$ after analysis of the 4-D Hough-Radon transforms. Using this information, we construct the 2-D HRTs by varying parametesr a, d. We choose the following bounds and quantization: $a \in [10 : 1 : 110]$ and $d \in [120 : 10 : 3500]$. The peculiar lower bound on d is to ensure that we do not exceed the boundaries of our spectrogram during integration. Note: With our prior knowledge of the fundamental frequency (213 Hz), we could have made more intelligent choices for the quantization of d. However, we choose rather to illustrate the HRT in all of its brute-force glory.

The 2-D Hough-Radon transform of the spectrogram is shown in Fig. 21. The high intensity regions in the HRT correspond precisely with the estimated a, d parameters of the vibrato. Figure 22 shows the histogram and cumulative histograms of the HRT. The results of thresholding at $p = 0.80$ and $p = 0.99$ are shown in Fig. 23.

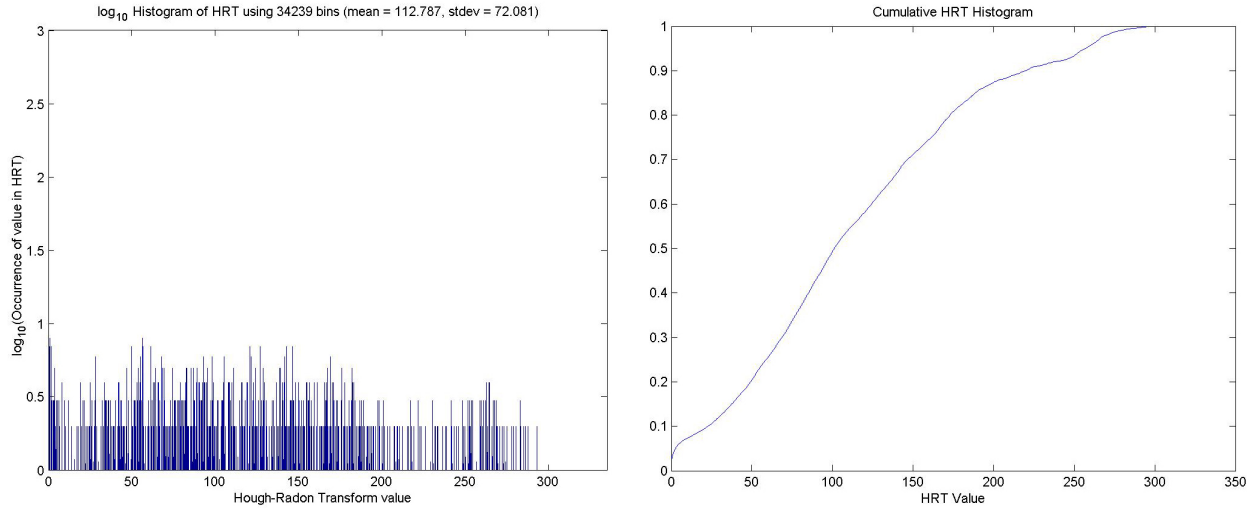


(a) (b)

Figure 21: 2-D HRT of spectrogram in Fig. 18: (a) linear scale, (b) $\log_{10}$ scale

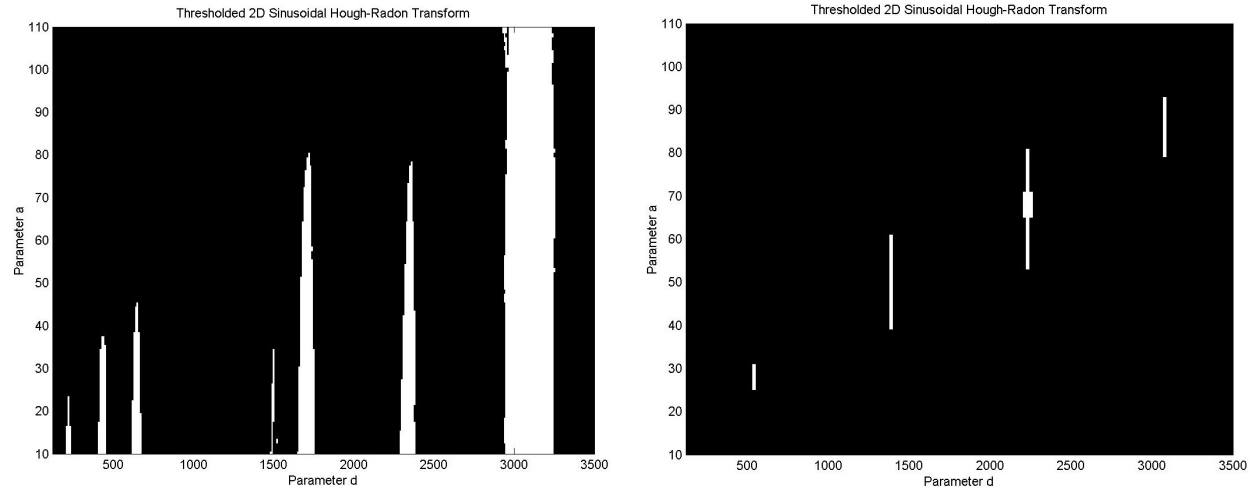
5 Future Work

We suggest the following ideas for improvement the program.



(a) (b)

Figure 22: (a) Histogram and (b) cumulative histogram of 2-D HRT of spectrogram in Fig. 18



(a) (b)

Figure 23: Thresholded HRTs of spectrogram in Fig. 18: (a) $p = 0.80$, (b) $p = 0.99$

5.1 Stockwell Transform

We suggest using the Stockwell Transform (S-transform) to generate the time-frequency representation of the audio signal [20]. The S-transform is a particular continuous wavelet transform that that uses the Morlet wavelet. For a time series signal $u(t)$, it is defined as

$$S(\tau, f) = \int_{-\infty}^{\infty} u(t) w(\tau - t, f) e^{-i2\pi f t}\, dt, \tag{7}$$

where the windowing function $w(\tau - t, f)$ is a Gaussian centred at time τ:

$$w(\tau - t, f) = \frac{|f|}{k\sqrt{2\pi}} e^{\frac{-f^2(\tau - t)^2}{2k^2}}. \tag{8}$$

The parameter k is used to scale the width of the window. The S-transform is similar to the STFT, except that the windowing function w is automatically scaled with frequency. Narrower windows are used to localize high frequency content, while wider windows are used to detect lower frequency variations. This operation is intuitive, as low frequency variations in a signal take place over longer time duration than high frequency variations. Since the window is not of a fixed width, this scaling means that more optimal time-frequency localization can be achieved than with the STFT. In addition, it is not necessary to search for an optimal window width.

5.2 Time-Frequency Reassignment

As mentioned in section 2.2, localization of structures could be significantly improved using a reassignment method[14, 15, 16]. From our experience, these methods effectively localize most synthetically generated frequency modulated signals in the spectrogram.

5.3 Detection of Incomplete Sinusoids

We refer to "incomplete" sinusoids as those for which any of the parameters a, b, c, d are not constant within the analysis time frame, as in Fig. 2. It is apparent that our current methods fail to conclusively detect incomplete sinusoids. These may appear in the spectrogram for a number of reasons: changing of pitch, volume, nasality, breathing, etc.

We propose to search for such incomplete sinusoids by adding a time accumulator to each bin of the HRT. Thus, $B(a, b, c, d)$ becomes $B(a, b, c, d, t)$. For given parameters (a_i, b_i, c_i, d_i), the univariate function $B(a_i, b_i, c_i, d_i, t),\ t = 0, \ldots, t_{\max}$, is the cumulative value of the bin value up to time t:

$$B(a_i, b_i, c_i, d_i, t) = \int_{\tau=0}^{\tau=t} S(\tau, \omega_{(a,b,c,d)}(\tau))\, d\tau, \tag{9}$$

where S is the spectrogram. For bins with relatively high values $B(a_i, b_i, c_i, d_i, t_{\max})$, a discontinuty in $B(a_i, b_i, c_i, d_i, t)$ at $t = t'$ indicates the sudden appearance of a sinusoid with parameters (a_i, b_i, c_i, d_i) at time t'. Likewise, if the function $B(a_i, b_i, c_i, d_i, t)$ levels out (has low slope) between $t = t'$ and $t = t''$, then we are alerted to the disappearance of the sinusoid between t' and t''. This analysis will most likely require computation of the time derivatives $\frac{d}{dt}B(a_i, b_i, c_i, d_i, t)$.

5.4 Parameter Optimization Strategy

It is clear from the discussion in section 3.2 that searching for sinusoid parameters is akin to finding maximizers in the H-R domain. Thus, we propose to use a numerical optimization strategy (e.g. Nelder-Mead Simplex gradient ascent or Powell's method [21]) to determine these maxima. The function to be maximized is the 4-D HRT bin value $B(a, b, c, d)$. The numerical optimizer would decide which parameters to pass to the HRT in order to compute the bin values, rather than blindly computing $B(a, b, c, d)$ for all parameters within the search space. We note that this optimization will be complicated by the fact that the H-R domain has many local minima and maxima. Thus, both local and global optimizations may be in order.

References

[1] D.L. Jones. Understanding vibrato. *http://www.voiceteacher.com/vibrato.html*, 2006.

[2] E. Prame. Measurement of the vibrato rate of ten singers. *J Acoust Soc Am*, 96(4):1979–1984, 1994.

[3] C.R. Nave. Sundberg's singing formant. *http://hyperphysics.phy-astr.gsu.edu/hbase/music/singfor.html*, 1999.

[4] R. Miller. *The structure of singing: system and art in vocal technique*. Schirmer Books, 1996.

[5] P.H. Dejonckere, M. Hirano, and J. Sundberg. *Vibrato*, chapter Acoustic and Psychoacoustic Aspects of Vocal Vibrato, pages 35–62. Singular Publishing Group, 1995.

[6] C. Leydon, J.J. Bauer, and C.R. Larson. The role of auditory feedback in sustaining vocal vibrato. *J Acoust Soc Am*, 114(3):1575–1581, 2003.

[7] J. Sundberg. *The Science of the Singing Voice*. Northern Illinois University Press, 1987.

[8] E. Prame. Vibrato extent and intonation in professional western lyric singing. *J Acoust Soc Am*, 102(1):616–621, 1997.

[9] S. Wood. What are formants? *http://www.ling.lu.se/persons/Sidney/praate/whatform.html*, January 2005.

[10] L. Cohen. Time-frequency distributions: a review. *Proc IEEE*, 77:941–981, 1989.

[11] L. Cohen. *Time Frequency Analysis: Theory and Applications*. Prentice Hall, 1994.

[12] K. Gröchenig. *Foundations of Time-Frequency Analysis*. Birkhäuser Boston, 2000.

[13] R.M. Rangayyan. *Biomedical Signal Analysis*. John Wiley & Sons, 2002.

[14] F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Trans Sig Proc*, 43(5):1068–1089, 1995.

[15] S.A. Fulop and K. Fitz. A spectrogram for the twenty-first century. *Acoustics Today*, pages 26–33, 2006.

[16] F. Plante, G. Meyer, and W.A. Ainsworth. Improvement of speech spectrogram accuracy by the method of reassignment. *IEEE Trans Speech Audio Proc*, 6(3):282–286, 1986.

[17] J. Canny. A computational approach to edge detection. *IEEE TPAMI*, 8(6):670–698, 1986.

[18] R.M. Rangayyan and S. Krishnan. Feature identification in the time-frequency plane by using the hough-radon transform. *Pattern Recognition*, 34:1147–1158, 2001.

[19] R.M. Rangayyan. *Biomedical Image Analysis*. CRC Press, 2005.

[20] R.G. Stockwell, L. Mansinha, and R.P. Lowe. Localization of the complex spectrum: The S transform. *IEEE Trans Sig Proc*, 44(4):998–1001, 1996.

[21] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 2002.