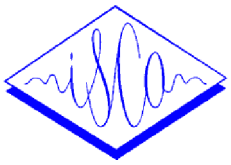

Models and Analysis of Vocal
Emissions for Biomedical Applications
(MAVEBA 1999)
Firenze, Italy, September 1-3, 1999

A HIGH-RESOLUTION TIME-FREQUENCY ANALYSIS OF THE SINGING VOICE

Maureen Mellody
Applied Physics Program, 500 East University Dr.,
University of Michigan, Ann Arbor, MI 48109-1120
mmellody@umich.edu

Gregory H. Wakefield
Department of Electrical Engineering and Computer Science, 1301 Beal Ave.,
University of Michigan, Ann Arbor, MI 48109-2122
ghw@eecs.umich.edu

We apply the modal distribution, a high-resolution time-frequency distribution, to the study of sung musical passages. Evidence is presented comparing the modal distribution with the spectrogram for a set of synthetic signals which emulate human singing. We then compare the two techniques for sung passages from four student sopranos with respect to measures of the instantaneous frequency and amplitude of the partials. In general, the modal distribution appears to be a more sensitive measure of individual differences in vocal production than the spectrogram.

1. INTRODUCTION

The spectrogram is a tool that is commonly used to extract time and frequency information from a vocal signal. This technique can capture many salient features of a sung passage, such as pitch and formant structure. However, due to the spectrogram's inherent limitations in trading temporal with spectral resolution, many of the finer temporal variations in the partials of a vocal signal can be lost. These variations appear to be important in the perception of the sung passage, as their absence in a synthesized signal often produces an artificial quality and, in the worst case, a loss of singer identity.

Time-frequency distributions have long been used by physicists and communications engineers, and have recently become a research focus in signal processing [1]. One general set of time-frequency distributions is Cohen's class [2], in which each member can be represented as the convolution of the Wigner distribution (e.g., the Fourier transform of the signal's instantaneous autocorrelation function) with a two-dimensional kernel function. Depending on the choice of kernel, time-frequency distributions trade temporal and spectral resolution against a nonlinear component in the transform in which "energy" in the signal's distribution can appear at frequencies and times when no such energy actually exists in the signal (the so-called cross-terms of the distribution). The spectrogram, in particular, belongs to Cohen's class, and utilizes a kernel that substantially suppresses crossterms at the expense of limiting simultaneous time and frequency resolution. An alternative to the spectrogram, the modal distribution [3], utilizes a kernel that can still suppress cross-terms by more than 50 dB but significantly improves the trading ratio of time and frequency resolution over that provided by the spectrogram. The modal kernel was designed explicitly for signals that can be represented by the real part of a series of complex exponentials, each with slowly-varying amplitude and phase. This signal model applies to many musical signals, including those of the sung voice.

In this work we apply both representations, the modal distribution and the spectrogram, to develop a high-resolution portrait of sung musical passages. Both representations reveal the partials of a sung passage as isolated ridges in the time-frequency image. For each partial, the instantaneous frequency and amplitude at each analyzed time interval is determined by the centroid of each isolated ridge and the square root of the summed energy across the bandwidth, respectively.

In Section 2, we discuss results from the analysis of a set of test signals designed to mimic properties of vocal signals. We compare the results obtained from the spectrogram and

the modal distribution for these signals with known instantaneous amplitude and frequency. In Section 3, we then apply each technique to the study of sung passages recorded from four female vocalists.

2. EVALUATION OF DISTRIBUTION PERFORMANCE FOR SYNTHETIC SIGNALS

There are two ways to compare the performance of the modal distribution and the spectrogram when applied to singing. The first uses a mathematical criterion: how do the extracted amplitude and frequency values compare to the true values? A second method of evaluation uses a perceptual criterion: does a signal synthesized from the extracted estimates sound like the original signal? We first apply the spectrogram and the modal distribution to a series of test signals with components of known instantaneous amplitude and frequency. We use numeric criteria to determine the success of the analysis by comparing the extracted estimates to the actual values used to create the signal.

The modal kernel was initially designed for signals modeled as the real part of a linear combination of sinusoids, each with amplitude and frequency that do not vary with time, i.e. the amplitude and frequency are constant for each partial. Estimates based on the modal distribution have been shown, both in simulation as well as in application, to be superior to those of the spectrogram, given certain classes of time-varying acoustic signals [4-6]. This includes amplitude variation that is well-described by an exponential decay (such as piano tones), periodic frequency variation on the order of a quarter of a semitone (such as violin vibrato), and amplitude modulation with a modulation index up to 1 (such as flutter-tongue flute and violin vibrato). We extend these observations to include simulated vocal vibrato passages, in which the frequency excursions can exceed a semitone in variation.

2.1. Analysis Parameters

In its analytic form, a modal distribution image can be computed with time and frequency axes that are continuous variables. However, to compute the modal distribution image for a waveform digitized at 44.1 kHz, it is necessary to create a discrete image, where each axis has an associated sampling interval. For the modal distribution of the signals in this work, the update interval along the time axis was chosen to be 5 ms. This interval was chosen so that each vibrato cycle in the singer's note is represented by on the order of 40 time slices. Moving to a shorter time interval did not significantly improve performance and only added to the computational requirements of the transform; variations on time scales shorter than 5 ms were apparently not meaningful to the perceived quality. The analysis interval along the frequency axis was chosen to be 21.5 Hz. In general, this frequency interval is chosen to guarantee that each partial is isolated in frequency in the time-frequency image. It is also chosen to give a transformation length that is a power of two to utilize a Fast Fourier Transform in the calculation. It should be noted that the choice of time interval defines the resolution in time (5 ms) of the partial's time variation but that the resolution of the partial's frequency estimate is NOT determined by the frequency interval chosen. We receive improved frequency resolution due to the centroid calculation applied to each ridge in the time-frequency surface.

For the spectrogram, the Hamming window length and the amount of overlap were chosen to acquire a time-frequency matrix of size equal to that of the modal distribution, i.e. with a time bin size of 5 ms and a frequency bin size of 21.5 Hz. The centroid operators described above are applied to the time-frequency surfaces of both representations.

2.2. Simulated Vibrato: Partials with Sinusoidal Frequency Modulation and Constant Amplitude

Table 1 summarizes the estimates of five synthetic signals designed to simulate the frequency variations seen in vocal vibrato. The test signals all consisted of a fundamental at A4 (440 Hz) and its first seventeen overtones. Sinusoidal frequency modulation was added with a rate of 6 Hz and an excursion of cents. These values approximate the rates and excursions reported for soprano singers. For each partial, the amplitude was constant. The relative amplitude values of the partials were chosen to simulate each of the five primary vowels. The results in Table 1 show the worst-case deviation from the actual values in cents, across all partials for each note. The error in the

amplitude estimates are shown in dB relative to the actual value.

	Spectrogram		Modal	
Note	Freq (cent)	Amp (dB)	Freq (cent)	Amp (dB)
/ɛ/	5.87	-10.8	4.86	-28.8
/i/	6.16	-12.2	4.23	-29.3
/a/	5.92	-10.1	3.94	-29.5
/o/	6.17	-11.1	2.75	-31.1
/u/	6.82	-12.5	4.79	-30.3

Table 1: Largest deviations in estimates for sinusoidal FM and constant amplitude A4 test signals.

In all cases, the errors in frequency and in amplitude were smaller for the modal distribution than the spectrogram, particularly for the amplitude estimates. The worst-case results occurred at extrema in the frequency modulation for all cases above. The data in Figure 1 suggests that worst-case errors reported for the modal distribution are significantly larger than the average error. Figure 1 reports the worst errors on a partial-by-partial basis for a simulated /a/ note. For the stronger partials, the amplitude errors are close to 90 dB down from the maximum value, and the frequency errors are well below a cent. In the figure, the upper panel shows worst errors in frequency (reported in cents), and the lower panel shows worst errors in amplitude (reported in dB relative to the actual value). In the lower panel, then, the longer bars (those which extend to the most negative value) indicate better performance, whereas smaller bars in the upper panel indicate better performance.

For the spectrogram, the worst-case errors more closely resemble the average error for each partial; a plot similar to Figure 1 for the spectrogram would show bars of essentially the same magnitude within each panel of the figure.

In general, the errors appear to increase on the partials with the lowest overall strength, such as the ninth partial and above in the above example. Figure 2 shows the strengths of the partials, shown in relative dB. The worst estimates tend to occur on low-strength partials in the neighborhood of a higher-gain partial. There is a drop in strength between the eighth and ninth partials of nearly 30 dB in this /a/ note, and the worst estimates are correspondingly on partial nine.

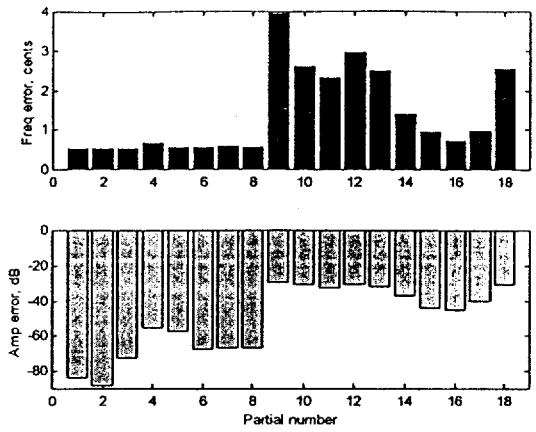

Figure 1: Errors in the modal distribution estimates by partial for the simulated vibrato /a/ note with frequency modulation and constant amplitude. The top panel shows worst error in frequency (in cents) and the bottom panel shows worst amplitude errors (in dB down from the actual value of the partial).

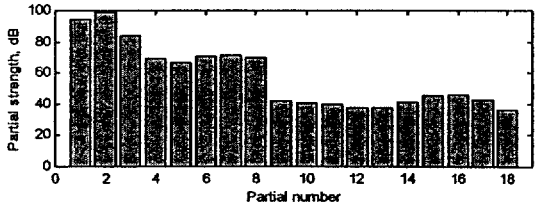

Figure 2: Relative partial amplitude strength as a function of partial number, shown in dB, for the simulated /a/ note.

2.3. Simulated Vibrato: Partials with Sinusoidal Frequency Modulation and Sinusoidal Amplitude Modulation

Table 2 summarizes the estimates of synthetic signals with both frequency and amplitude modulation, which is a more realistic representation of vocal vibrato. In these test signals, the frequency and average amplitude values used were the same as in the previous section. Sinusoidal amplitude modulation was added at a rate of 6 Hz (equal to the frequency modulation), with the amplitude modulation in phase with the frequency modulation. The modulation depth was 0.67 on all partials, which is larger than the modulation observed on partials from voice signals. Again, the values shown in Table 2 represent the largest deviations from the actual values.

The positive values in the dB error for the spectrogram indicate that the error is at least as large as the value of the signal at that point. These instances occur at the minima in the amplitude modulation, where the signal value is quite small. The errors in frequency are

equivalent to the errors shown in Figure 1 for test signals without amplitude modulation. The amplitude estimates for the modal distribution degrade only slightly, but the amplitude errors in the spectrogram estimates increase significantly.

	Spectrogram		Modal	
Note	Freq (cent)	Amp (dB)	Freq (cent)	Amp (dB)
/ɛ/	7.27	1.95	5.34	-24.5
/i/	7.57	1.69	4.29	-25.1
/a/	7.09	1.94	3.24	-27.1
/o/	7.39	1.41	2.98	-27.1
/u/	8.35	1.11	5.37	-26.3

Table 2: Largest deviations in estimates for sinusoidal FM and AM test signals

Figure 3 shows the errors in frequency and amplitude estimation as a function of partial number for the simulated vibrato /a/ with sinusoidal modulation in both frequency and amplitude. Again, the worst estimates occur on the ninth partial. There still exists a "jump" in amplitude error between partials eight to nine, but the error change is roughly 20 dB rather than the 40 dB observed in the previous case.

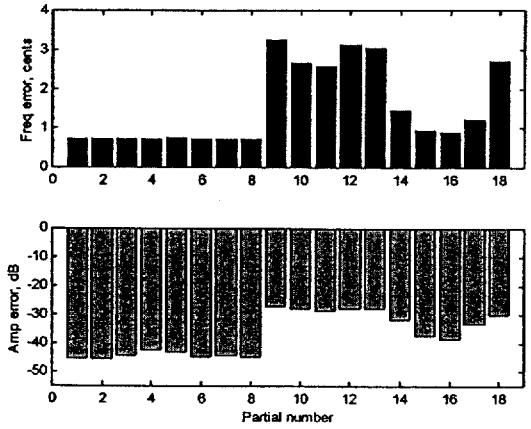

Figure 3: Errors in the modal distribution estimates by partial for the simulated vibrato /a/ note with both frequency modulation and amplitude modulation. The top panel indicates worst error in frequency (in cents) and the bottom panel indicates worst amplitude errors (in dB down from the actual value of the partial).

Test signals generated at other pitches with comparable FM rates and excursions provide similar results to the examples shown. Other choices of spectrogram parameters also resulted in comparable errors in the estimates.

3. SIGNAL ANALYSIS: SUNG PASSAGES

While analysis of synthetic signals indicates that the modal distribution provides a more accurate representation of the amplitude and frequency of the note's partials, this improvement in resolution only holds for real voice recordings if those recorded signals obey the constraints of our signal model, i.e. that the partials are isolated in frequency with slowly-varying amplitude and frequency values. To evaluate the performance of the spectrogram and the modal distribution when applied to sung passages, a perceptual criterion is used, as there are no "actual" values for a mathematical comparison. We incorporate the estimates extracted from each time-frequency distribution into an additive synthesis model. The re-synthesized signal is created from a sum of sinusoids, each with time-varying amplitude and frequency as determined from the analysis. The estimates of amplitude and frequency are linearly interpolated between each analysis interval to re-create a sound sampled at the original recording's sample rate. These re-synthesized signals (one from each representation) and the original recording are compared aurally to determine the degree to which the perceptually salient features are captured.

In addition, the re-synthesized notes are also subjected to the same analysis methods, giving the time-varying amplitude and frequency content of the re-synthesized signal. In this way, we are not required to rely on perceptual validation but can numerically determine the degradation in estimate accuracy caused by the analysis tool.

3.1. Signal Acquisition and Analysis Parameters

We analyzed signals recorded from four sopranos, all students in the voice program at the University of Michigan. Each soprano sang the same musical passage, a descending five-note scale beginning on D5 with the vowel sequence /ɛ/-/i/-/a/-/o/-/u/. The wave files were analyzed with the modal distribution at 5 ms intervals and a transform length of 1024 points. Analogous parameters for the spectrogram were used.

3.2. Analysis of Vocal Vibrato

Figure 4 shows the time-varying fundamental frequency extracted from the modal distribution. Each panel contains the fundamental contour from a different singer. The degree of vibrato ranged from very little in the

case of S1 to a continuous vibrato in the case of S4. The vibrato rates range from roughly 5 Hz to 6.5 Hz.

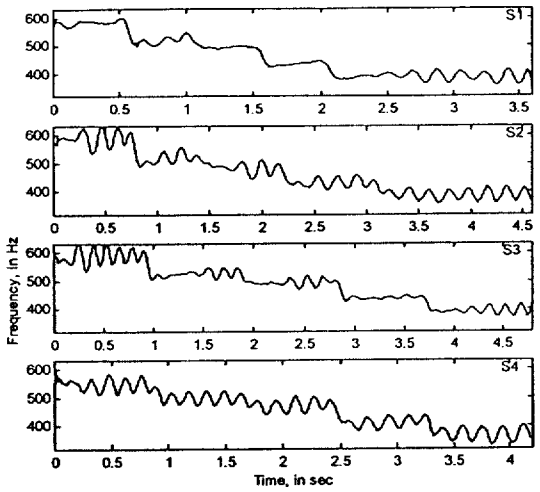

Figure 4: Fundamental frequency estimates as a function of time extracted from the modal distribution for four soprano singers. The passage sung is a descending five-note scale on the primary vowel series.

There is little information to be gained by displaying the frequency estimates extracted from the spectrogram analysis of the same musical passages, as the features observed visually are quite similar when viewed over the duration of the full phrase as in Figure 4.

3.3. Comparison of Spectrogram and Modal Distribution Methods for Sung Passages

When signals are re-synthesized by incorporating the estimates into a sum of sinusoids, the differences between the two methods becomes readily apparent when compared aurally. In the case of the spectrogram, the resulting synthesis sounds unnatural and the singer identity is somewhat obscured, while the modal distribution synthesis sounds essentially identical to the original recording. Examples of this can be found on our web site, at http://musen.engin.umich.edu/.

While the long time-scale view of the estimates as in Figure 4 reveal no major distinctions between the two representations, a short time-scale view shows differences in fine temporal features. This is illustrated in Figure 5. The signal shown is the third partial of the /i/ portion of the sound sung by singer S3. The amplitude estimate based upon the modal distribution analysis is shown as a dotted line for a half-second portion of sound. The solid line shows the spectrogram estimate from the original signal.

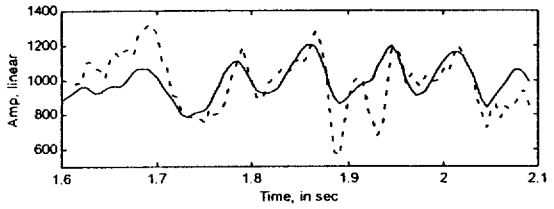

Figure 5: Amplitude in linear units as a function of time for the third partial of /i/ sung by singer S3. The dashed line is the modal distribution estimate of the original signal. The solid line is the spectrogram estimate of the original signal.

The two methods, modal and spectrogram, provide slightly different results for the original signal; however, since we have no exact knowledge of the temporal behavior of the original signal, we cannot say which is more accurate. We presume that the spectrogram tends to smooth the fine time-scale features of the signal and the modal distribution tends to add artifact to the signal. To determine the extent of these possible deteriorating factors, we can re-synthesize signals from both estimates and analyze the resulting waveforms.

Figure 6 shows the results of analyzing the signal synthesized from the modal distribution analysis. In both panels, the modal distribution values are shown as a dashed line. The modal distribution estimates of this modal-synthesized signal are shown in the top panel with a solid line. In the bottom panel, the solid line shows the spectrogram estimates of the modal-synthesized signal. The modal distribution estimates of the modal-synthesized sound are essentially identical to the input values. The spectrogram, in the lower panel, shows a more smoothed result, similar to the spectrogram estimates of the original signal in Figure 5.

In the final figure, we now analyze the signal re-synthesized from the spectrogram estimates of the original signal. The spectrogram estimate is shown in each panel as a dashed line. The solid line in the top panel shows the modal estimates of this signal, while the lower panel shows the spectrogram estimates of this signal. The modal distribution estimates so closely resemble the input signal that the dashed line is not visible, as it lies beneath the solid line. When analyzing the spectrogram-synthesized note, the modal distribution again produces little, if any,

distortion to the estimated values. The spectrogram, on the other hand, produces estimates that are smoothed even further from the original spectrogram-based analysis.

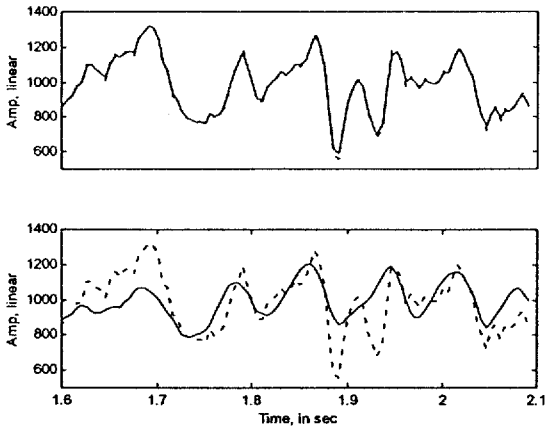

Figure 6: The dashed line in both panels shows the modal distribution estimate of the original signal for the third partial of a portion of the /i/ vowel for singer S3. The solid line in the top panel shows the modal distribution estimates of the modal-synthesized sound, and the solid line shows the spectrogram estimate of the modal-synthesized sound.

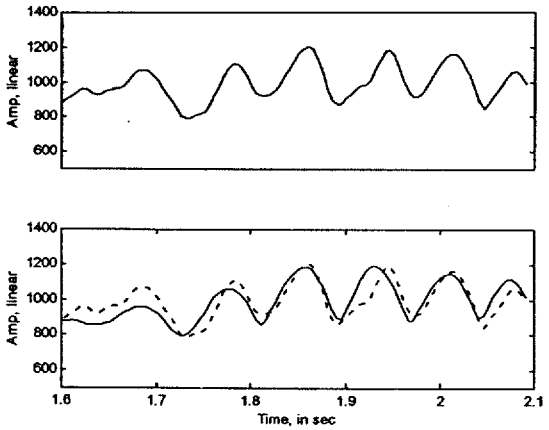

Figure 7: The dashed line in both panels shows the spectrogram estimate of the original signal for the third partial of a portion of the /i/ vowel for singer S3. The solid line in the top panel shows the modal distribution estimates of the spectrogram-synthesized sound, and the solid line shows the spectrogram estimate of the spectrogram-synthesized sound.

In these examples, the modal distribution analysis provides estimates that more closely approximate the original values than the spectrogram. This is true in cases of smoothly-varying functions, such as those in Figure 7, or functions with short time-scale features, such as those in Figure 6. The spectrogram, on the other hand, tends to obscure these time-varying details.

4. CONCLUSIONS

In both the synthetic examples and recorded sung passages, the modal distribution exhibits superior performance over the spectrogram when compared with either mathematical or perceptual criteria. Signals re-synthesized from the modal distribution estimates are virtually indistinguishable from the original recordings in the case of isolated sung vowels. Using this improved resolution, we are able to resolve fine-scale temporal variations in sung signals that have previously been undetected by spectrogram-based methods.

5. REFERENCES

[1] Cohen, L. (1966). "Generalized Phase-Space Distribution Functions," *J. of Math Phys.*, 7, 781-786.

[2] Cohen, L. (1995).*Time-Frequency Analysis* (Prentice Hall: Englewood Cliffs, NJ).

[3] Pielemeier, W.J. and Wakefield, G.H. (1996). "A High-Resolution Time-Frequency Representation for Musical Instrument Signals," *J. Acoust. Soc. Am.*, 99(4), 2382-2396.

[4] Guevara, R.C.L. (1997). *Modal Distribution Analysis and Sum of Sinusoids Synthesis of Piano Tones,* Ph.D. Thesis, University of Michigan, Ann Arbor, Michigan.

[5] Mellody, M. and Wakefield, G. H. (1997). "A Modal Distribution Study of Violin Vibrato," *Proc. of Intl. Comp. Music Conf.*, Thess., Greece, Sept. 1997.

[6] Pielemeier, W.J. and Wakefield, G.H. (1994). "Multi-Component Power and Frequency Estimation for a Discrete TFD," *Proc. of IEEE Symp. on Time-Freq. and Time-Scale Anal.*, Philadelphia, PA, Oct. 1994.