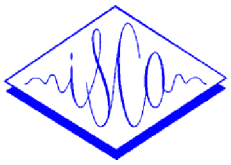

Models and Analysis of Vocal
Emissions for Biomedical Applications
(MAVEBA 1999)
Firenze, Italy, September 1-3, 1999

CHROMAGRAM VISUALIZATION OF THE SINGING VOICE

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

The chromagram is a transformation of a signal's time-frequency properties into a temporally varying precursor of pitch. This transformation is based on perceptual observations concerning the auditory system and has been shown to possess several interesting mathematical properties. We illustrate the use of the chromagram as a fast and robust method for visualizing attributes of a singer's voice, which is relatively invariant to changes in the vocal tract resonances.

1. INTRODUCTION

Voice science has used the spectrogram to visualize and study the acoustics of the vocal signal. Although a useful representation of the signal, the spectrogram, as a visual tool, can be faulted because pitch and resonance in vocal production are not readily separable in the representation. For example, in a narrowband spectrogram, we draw inferences about the pitch of the glottal source by looking for equal spacing among the "spectral ridges" in the display. The resonant properties of the vocal tract are inferred from variations in the relative amplitudes of such ridges. In this paper, we introduce an alternative representation that can be derived from the spectrogram and provides a more direct measure of those glottal source variations related to pitch.

2. MAPPING ONE DIMENSION INTO TWO: AN AUDITORY REPRESENTATION OF FREQUENCY

In the early 1960's, Shepard reported that two dimensions rather than one are necessary to represent the perceptual structure of pitch [1]. He determined that the human auditory system's perception of pitch was better represented as a helix, than as a straight line, and coined the terms *tone height* and *chroma* to characterize the vertical and angular dimensions, respectively. In this representation, as the pitch of a musical note increases, say from C1 to C2, its locus moves along the helix, rotating chromatically through all of the pitch classes before it returns to the initial pitch class (C) one cycle above the starting point.

According to Shepard's results, the pitch (p) of a signal can be further divided into values of chroma (c) and tone height (h)

$$p = 2^{c+h}$$

For this decomposition to be unique, it is sufficient for chroma to be a real number between *0* and *1*, and for tone height to be an integer. The implication of Shepard's representation is that the distance between two pitches depends on c and h, rather than on p alone.

With respect to music, Shepard's findings are not surprising and instead, confirm the centuries-long common practice of giving special emphasis to octave relations among notes. Indeed, music theorists use the terms *pitch class* and *octave number* to denote chroma and tone height, respectively. We will continue to use Shepard's terms, however, because of a more radical interpretation of Shepard's work that was advanced in the mid 1980's.

In the process of developing his Auditory Image Model in computational audition, Patterson generalized Shepard's results to *frequency*. Even though he substituted the Archimedian spiral for Shepard's helix as a basic representation of frequency in the auditory system, the mapping from one dimension to two remains effectively the same. His pulse-ribbon model [2] transforms each temporal frame of the auditory image into an activity pattern along a spiral of temporal lags, such that lag values along the same "spoke" of the spiral are octave multiples of each other.

Structurally, the frequency dual of a spiral decomposition using Patterson's lag variables is

equivalent to the pitch structure of Shepard's, e.g.,

$$f = 2^{c+h}$$

where chroma denotes the angle along the spiral and tone height denotes the position along the common spoke. As in the case of pitch, a sufficient condition for this decomposition to be unique is that $c \in [0,1)$ and $h \in Z$, the set of integers. Similar to ideas of pitch, certain frequencies under this system share the same *chroma class* if and only if they are mapped to the same value of c. Thus, 200, 400, and 800 Hz share the same chroma class as 100 Hz, but 300 Hz does not.

Such a mapping of a one-dimensional variable, like frequency or time lag, onto a two-dimensional coordinate system appears, at first glance, to increase, rather than decrease, the complexity of an acoustic signal's representation. However, it implies a different type of geometry in the signal representation, which may have its own unique advantages.

Our work has studied the mathematical consequences of such two-dimensional representations of frequency as chroma and tone-height, particularly with respect to the estimation of pitch; first, in the case of pitch contours of Putonghua Chinese [3], and subsequently with respect to pitch processing in single- and multi-voiced music signals [4,5]. Throughout this research, we have observed that chroma is not only useful as an intermediate representation in the formation of pitch estimate, but as a representation, in its own right, of the acoustic signal.

One important result is the nature of the mapping between the frequencies and chroma of a harmonic series. Thus, for a given fundamental f_0, the 1st, 2nd, 4th, 8th, 16th, etc. partials belong to the same chroma class; the 3rd, 6th, 12th, etc. belong to another; and so on, such that the chroma of each odd harmonic forms a chroma class for all octaves of that harmonic. This concentration of harmonics around a smaller set of chroma has proven to be a useful property in extracting pitch from signals and in simplifying the visual representation of harmonic signals as they vary in pitch over time.

3. The Chromagram

The chromagram *s(t,c)* extends the concept of chroma to include the dimension of time. Just as we use the spectrogram *s(t,f)* to infer properties about the distribution of a signal's energy over frequency and time, the chromagram can be used to infer properties about the distribution of a signal's energy over chroma and time.

Two issues arise in developing the concept of a chromagram. The first is one of definition. Chroma is a many-to-one mapping of frequency to the [0,1) interval while the chromagram is a measure of *signal strength* as a function of chroma and time. Therefore, the chromagram is a many-to-one mapping of signal strength at frequencies belonging to the same chroma class, e.g.,

$$s(t,c) = G\left(s(t,f); \forall f = 2^{c+h}\right)$$

The second issue concerns how to compute the chromagram. We have investigated several approaches. These can be divided between direct transforms of the original signal and transforms of a time-frequency image, e.g., the spectrogram, of that signal. In the following, we present results for the latter case. As we shall see, given the practical constraint of forming a signal's time-frequency image along a lattice of points, the form of the many-to-one function G becomes moot.

4. A FAST ALGORITHM FOR COMPUTING THE CHROMAGRAM

We compute the chromagram in two stages.

Stage 1. Construction of *s(t,f)*. The first stage computes a time-frequency image of the signal $s(t)$, which localizes the distribution of energy in the signal as a function of time and frequency. The spectrogram, for example, is one such method that is known to be relatively poor in localizing time and frequency when compared with other bilinear time-frequency distributions [6], but is nevertheless computationally efficient. In what follows, we adopt the spectrogram as our measure of the signal's time-frequency image.

Stage 2. Construction of *s(t,c)* based on *s(t,f)*. In its most precise mathematical form, such a construction requires knowledge of the many-to-one mapping G. However, in practice, we always operate on discrete-time signals for which *s(t,f)* is sampled on a lattice of points in

the time-frequency plane. This lattice structure linearly samples the image in time (what we shall call the *update or frame rate*) and frequency (what we shall call the *frequency bin width*). For example, if we compute the spectrogram of an 8 kHz sampled signal using a 512-pt. FFT every 10 ms, the update rate is 100 frames per second and the frequency bin width is 15.63 Hz. The intrinsic problem with any direct mapping of time and frequency onto time and chroma is that the linear lattice in frequency becomes a nonlinear lattice in chroma. This nonlinear sampling introduces objectionable visual artifacts in the time-chroma images of the signal, which can be reduced by increasing the length of the FFT, but not entirely eliminated.

In the following, we introduce the concept of *frequency re-localization* and show how it can be used to construct accurate time-chroma images from a signal's time-frequency image. This results in two stages of chromagram processing: a re-localization stage of the signal's spectrogram followed by a nonlinear mapping from frequency to chroma.

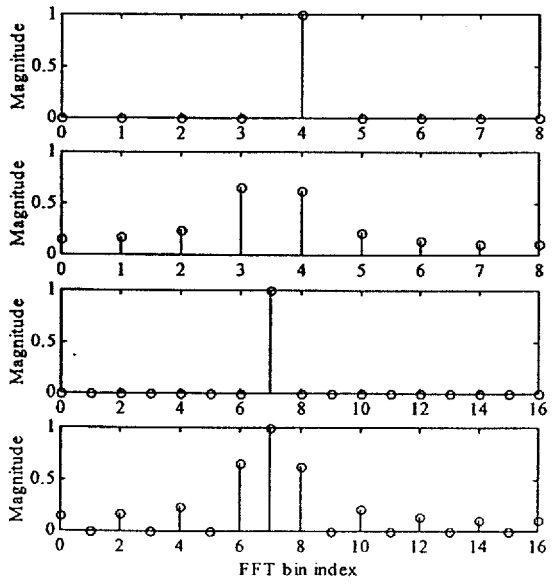

Figure One. The magnitude spectrum of a 16-point FFT of a sinusoid when the frequency of the sinusoid is one of the FFT lattice points is shown in the top panel. The lower three panels show magnitude spectra for a sinusoid halfway between the two lattice points of the first (top) panel for a 16-point signal/16-point FFT (second), a 32-point signal/32-point FFT (third), and a 16-point signal/32-point FFT (bottom).

Frequency Re-localization. Rather than further refinement of frequency bin width to reduce nonlinear sampling artifact in chroma, we propose a technique that re-localizes the sampling points of the original frequency bins according to properties of *s(t)*. To motivate this re-localization technique, consider the magnitude spectra of the FFTs in the two panels of Figure One.

The top panel shows the magnitude spectrum of the 16-point FFT of a 16-point sinusoid where the frequency of the sinusoid is one of the frequency lattice points of the FFT. The frequency of the sinusoid displayed in the lower three panels has been shifted from that in the top panel to lie halfway between the lattice points in the top panel. Each panel differs with respect to the lengths of the signal and FFT.

Given the spectrum in the top panel, we can conclude that the signal consists of a single sinusoid at the frequency indicated. When the frequency is shifted to lie halfway between lattice points in the FFT, it is more difficult to conclude that the signal consists of a single sinusoid. The second panel from the top shows the magnitude spectrum for a 16-point signal computed by a 16-point FFT. In this case, the evidence suggests a multi-component signal. Nevertheless, if we knew *a priori* that the signal consisted of a single sinusoid, we could use the fact that the magnitude spectrum in the bottom panel is symmetric about the "3.5" index to properly localize the frequency of that sinusoid.

In the present case, a simple method for recovering the "3.5" index is to refine the FFT lattice. If 32 samples of the signal are evaluated using a 32-point FFT, then the frequency is one of the lattice points and we can unambiguously conclude that the signal consists of a single sinusoid (third panel). However, if we are limited to a 16 point sample, as if often the case in spectrogram analysis where we are trading temporal and spectral resolution against each other, the 32-point FFT exhibits a maximum at the appropriate lattice point, but the profile is that of a multi-component signal as shown in the bottom panel.

The problem of frequency localization for signals dominated by a monochromatic component appears in several domains of signal processing. Spectral estimation theory has shown that the best localizer, in a maximum likelihood

sense, is the peak of the periodogram of a noisy sinusoid for a given amount of data. Several techniques in music analysis employ a "parabolic localizer" to refine the location of the maximum peak in the signal's magnitude spectrum. Such trackers best fit a parabola to three points of the FFT lattice – the peak and adjacent values above and below.

Our approach derives from methods proposed in time-frequency theory [6,7]. In this case, the centroid of a peak's neighborhood is used to refine the localization of that peak:

$$f_{ref}(t_k, n_{peak}) = \frac{\sum_{n\in\Omega} f_n s(t_k, f_n)}{\sum_{n\in\Omega} s(t_k, f_n)}$$

where the index set defines a neighborhood of the peak

$$\Omega = \{n_{peak} - L, \ldots, n_{peak}, \ldots n_{peak} + L\}$$

and n_{peak} is the index of the peak in the time-frequency image $s(t,f)$ evaluated on the lattice defined by the FFT length and update rate.

In the case of a monochromatic signal, the neighborhood of the peak in the time-frequency image can be extended to cover the entire frequency lattice for the given sample in time. The presence of additional sinusoidal components changes the shape of $s(t,f)$. To avoid contamination of the estimate from other sinusoidal components, we adopt a substantially smaller neighborhood. For example, we recommend an L of 3 for spectrograms based on 512 points for an 8 kHz sampling rate of the original signal, based on reasons given below.

Stage 2a. Refining the Time-Frequency Image. For the purposes of visualization, we apply the frequency re-localization procedure to the entire lattice of the signal's time-frequency image. This maps the image from a uniform to non-uniform lattice[1]:

$$s(t_k, f_n) \rightarrow s(t_k, f_{ref}(t_k, f_n))$$

Stage 2b. Mapping to Time-Chroma. Once the time-frequency image has been re-localized to a non-uniform lattice, we apply a much simpler mapping of $s(t,f)$ to the time-chroma plane

$$s(t_k, c_n) = s(t_k, c_{ref}(t_k, f_n))$$

by setting the multivariate function G to the identity operator and projecting each refined frequency to its chroma class

$$c_{ref}(t_k, f_n) = \log_2 f_{ref}(t_k, f_n) - \lfloor \log_2 f_{ref}(t_k, f_n) \rfloor$$

where $\lfloor\ \rfloor$ denotes the greatest integer function. Dropping the multivariate mapping function G appears to be warranted, given that in our experience the relatively sparse sampling of frequency yields refined frequencies with no common chroma class.

Specification of $s(t,f)$ and L. Stages 2a and 2b completely define the mapping procedure up to the definition of the time-frequency image of the signal and the width of the centroid neighborhood. These choices turn out to be integrally linked.

In our analysis, we have chosen to work with the spectrogram, a Fourier-based method common to speech research, rather than other nonlinear time-frequency distributions which may possess better time-frequency localization properties [8,9]. Of the parameters in the spectrogram, the choice of window, window length, and the length of the FFT interact with the choice of neighborhood in the frequency-refinement stage.

We have parametrically studied the problem of minimizing the refinement error over these design variables under the constraint that L should be reasonably small. The four panels of Figure Two show typical refinement errors for an optimal and sub-optimal solution given an FFT size of 512. The best of the standard windows is the 307-pt. Bartlett (top panel), which achieves a maximum error on the order of 0.0005 Hz over the lattice interval. This performance is an order of magnitude better than the Hamming window (next panel) of same length. The third panel shows the results of using a rectangular window in the spectrogram calculation, which still reduces the error by an order of magnitude when compared an estimate based on the lattice peak (bottom panel). In this latter case, lattice points are spaced every 15.6 Hz for an 8 kHz sampled signal evaluated by a 512-pt. FFT. Therefore, appropriate choice of frequency refinement can reduce the maximum error from 7.8 Hz by 4 orders of magnitude to 0.0005 Hz.

[1] Note that it is also possible to perform a similar calculation to refine each local energy estimate of the image. In the present application, this refinement does not lend any additional fidelity to the resulting visual image so we opt not to perform this step.

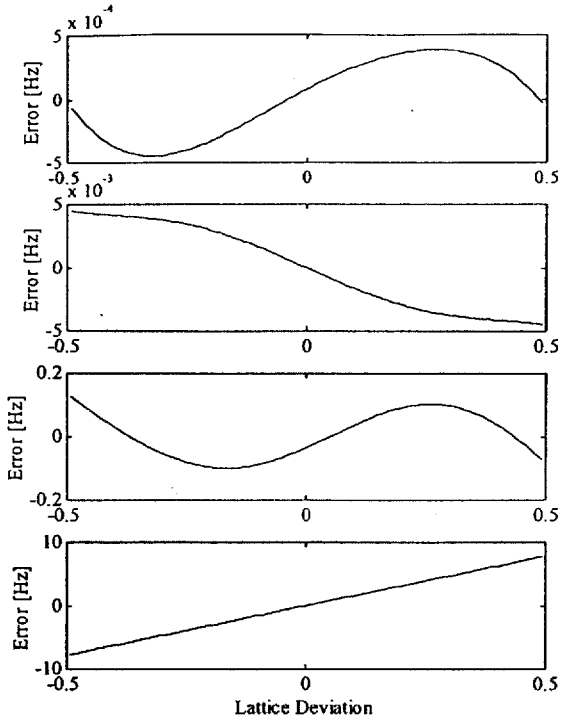

Figure 2. Error in frequency refinement is shown as a function of a sinusoid's frequency deviation from a lattice point of the 512-pt FFT. The four panels represent different window types for a fixed window length of 307. From top to bottom, the windows are Bartlett (best), Hanning, Rectangular, and no centroid calculation. The centroid was calculated over 7 points (L=3).

Such considerable reductions in error cost little computationally (8 multiplies and 14 additions per frequency of interest). If all 512/2 samples are refined (as we generally do in forming the time-chroma image), the entire process is computationally on the order of

$$N \log_2 N + 8N/2$$

In the case of 512, this evaluates to 6k multiplies. In comparison, brute force refinement by increasing the size of the FFT results in over 10k multiplies just to reduce the maximum error from 7.8 Hz to 3.9 Hz.

The reported errors are for the ideal condition of a pure monochromatic source. When additional monochromatic sources are present in the signal, we observe a similar rank ordering in the performance of each of the windows, but the error generally increases over that observed for a single sinusoid. For an equal amplitude partial series for a fundamental as low as 110 Hz, refinement errors remain small, but rapidly increase until at 55 Hz, the errors are on the same order as those observed if no refinement had been applied.

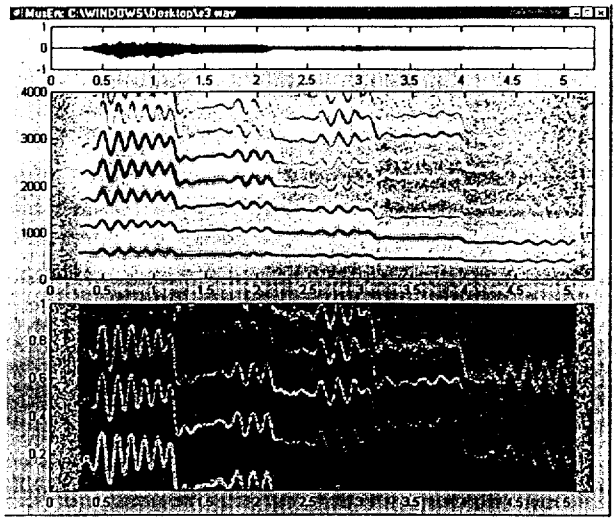

Figure 3. Three separate visualizations of a cardinal-vowel vocalise on a descending 5-note scale by a female student soprano are shown: time envelope (top), spectrogram (middle), and chromagram (bottom).

5. CHROMAGRAM VISUALIZATION OF THE SINGING VOICE

Using the two-stage process outlined in the previous section, we can visualize the chromagram of the singing voice. Figure 3 presents a cardinal-vowel vocalise on a descending 5-note scale by a female student soprano in the School of Music at the University of Michigan. The top panel represents the envelope of the temporal waveform whereas the middle and bottom panels represent the spectrogram and chromagram of the singer's voice, respectively. In both cases, time is presented horizontally, whereas frequency in Hz and chroma are shown vertically. The chromagram has been thresholded to a 15 dB dynamic range, whereas the spectrogram spans a 50 dB dynamic range.

In general, the spectrogram illustrates the vowel changes over the 5 seconds of the exercise through the variations. These are shown in the hot (0 dB) and cool regions (-50 dB) of each partial in the display (e.g., the 1-3 kHz range). Changes in pitch are more difficult to visualize in this display, however. While the descending nature of the pitch sequence can be observed in the variation of the fundamental frequency over time, many details are lost. One would be hard pressed to determined whether the singer was decreasing in half or whole steps, for example.

The chromagram complements the strengths of vocal tract visualization found in the

spectrogram. Variations in pitch are clearly demarcated as the singer transits through the octave equivalence from 1.2 to 2.2 seconds. Vibrato strength is also readily visualized. In the case of the spectrogram, the harmonic nature of the partials means that any vibrato will be magnified in its frequency excursion at relatively high partial frequency indices. Since chroma is organized logarithmically over the octave, one can visually divide the scale into twelve equal parts to denote semitone variations. This normalizes the vibrato excursions over the complete chroma range. Under this visualization, the growth of vibrato through the vowel is readily seen, as is the diminished excursion strength that occurs from the beginning to the end of the 5-second passage.

With respect to the goal of separating properties of vocal tract resonance from glottal source, we see that the chromagram's representation is dominated by properties of the glottal source, in contrast to the spectrogram, which is dominated by properties of the vocal tract resonances. Neither representation is a pure exemplar of either aspect of the singing voice. Further refinement of both could improve their respective domains of representation, but potentially at the same loss of robustness as observed in other signal processing approaches. Examples that we present in the oral version of this paper demonstrate the relatively insensitivity of chromagram visualization to the presence of orchestral accompaniment as well as the chromagram's sensitivity to variations in vocal production from pressed to breathy.

6. CONCLUSIONS

We introduced the chromagram as a method for visualizing properties of the singer's glottal pulse. Much of the paper developed the mathematical justification for our approach to calculating the chromagram. The outline of an algorithm was presented for implementing the calculations. The oral version of this paper focuses on several examples of singer production, based upon the analytic techniques presented in the written paper.

7. ACKNOWLEDGMENTS

This research was supported by a grant from the Ford Motor Company and by the MusEn Project at the University of Michigan with funding provided by the Office of the Vice President for Research. The author thanks Maureen Mellody, Bryan Pardo, and Profs. Freda Herseth and George Shirley for their contributions to this research.

8. REFERENCES

[1] Shepard, R. "Circularity in judgements of relative pitch," *J. Acoust. Soc. Am.*, vol. 36, pp. 2346-2353, 1964.

[2] Patterson, R. D. "Spiral detection of periodicity and the spiral form of musical scales," *Psychology of Music*, vol. 14, pp. 44-61, 1986.

[3] Duanmu, S., Wakefield, G. H., Hsu, Y., Qui, S., Guevara, R. C. "Taiwanese Putonghua Corpus (TWPTH) Speech and Transcripts," Linguistic Data Consortium, 1998.

[4] Wakefield, G. H. "The mathematical implications of a pulse-ribbon perceptual organization of pitch," Proc. of Intl. Comp. Music Conf., Hong Kong, August, 1996.

[5] Wakefield, G. H. "Time-Pitch Representations: Acoustical Signal Processing and Auditory Representations," IEEE Intl. Symposium on Time-Frequency Time-Scale, Pittsburgh, 1998.

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

[7] Pielemeier, W. J. and Wakefield, G. H. "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.

[8] Pielemeier, W. P. and Wakefield, G. H. "A high resolution time-frequency representation for musical instrument signals," J. Acoust. Soc. Am., Vol. 99(4), 2382-2396, 1996.

[9] Mellody, M. and Wakefield, G. H. "A High-Resolution Time-Frequency Analysis of the Singing Voice." Proceedings from this workshop, 1999.