CHARACTERIZING SINGING VOICE FUNDAMENTAL FREQUENCY TRAJECTORIES

Johanna C. Devaney[*]
McGill University
555 Sherbrooke Street West
Montreal, QC, Canada
devaney@music.mcgill.ca

Michael I. Mandel
Audience Inc.
440 Clyde Ave.
Mountain View, CA, USA
mmandel@audience.com

Ichiro Fujinaga[*]
McGill University
555 Sherbrooke Street West
Montreal, QC, Canada
ich@music.mcgill.ca

ABSTRACT

This paper evaluates the utility of the Discrete Cosine Transform (DCT) for characterizing singing voice fundamental frequency (F_0) trajectories. Specifically, it focuses on the use of the 1st and 2nd DCT coefficients as approximations of slope and curvature. It also considers the impact of vocal vibrato on the DCT calculations, including the influence of segmentation on the consistency of the reported DCT coefficient values. These characterizations are useful for describing similarities in the evolution of the fundamental frequency in different notes. Such descriptors can be applied in the areas of performance analysis and singing synthesis.

Index Terms**—** F_0 characterization, DCT, singing

1. INTRODUCTION

Fundamental frequency (F_0) trajectories in sung notes are characterized by instability at the beginning and ending of the note and by a relatively stable section of quasi-periodic pitch fluctuation (vibrato) in the middle. Previous studies of frequency-related information in the singing voice have focused on the characteristics of sung vibrato and the perceived pitch of each note, often defined as the mean of the vibrato [1]. Description of the perceived pitch of a note provides only a single value for each note and throws away information about how fundamental frequency (F_0) changes over the duration of the note. This paper explains how the evolution of F_0 can be characterized, specifically the slope and the curvature of the stable vibrato section of the note, with a normalized version of the Discrete Cosine Transform (DCT). Such a characterization is useful for both studying performance, and for generating more realistic sounding synthesized performances. The study of intonation singing performance is our primary goal, and we are particularly interested in exploring whether singers have different F_0 slopes and curvatures in different melodic contexts.

Once accurate F_0 estimates have been obtained, slope and curvature can be calculated in a number of different ways. One way of calculating slope is to take the difference between the first and the last value in a signal and divide it by its length. The problem with this approach is that it is highly sensitive to any noise in the first or last element. A more robust option is to calculate the slope and curvature over the duration of the signal rather than from just two points. One way of doing this is to fit a second-order polynomial to the signal in the least squares sense. The mean, slope, and curvature of a signal can be derived from such a second-order polynomial. In such a fit, however, the earlier terms of the polynomial estimate can change as the order of the polynomial increases. An orthogonal basis like the DCT [2] does not suffer from this problem. When approximating a function using DCT coefficients, the lower-order coefficients are independent of higher-order coefficients, and the "order" of the model does not affect the coefficient values. Another advantage of the DCT is that it provides a single value for slope and curvature estimates without the need to summarize the calculations over duration of the note. While other authors have looked at frequency analyses of F_0 trajectories, e.g., for the purposes of more realistically time-stretching speech and singing [3], our application of the DCT performs a post-processing normalization to convert the results to units (i.e., cents/second and cents/second^2) that are more relevant to the task of comparing different recorded F_0 trajectories than pure frequency analysis.

2. DISCRETE COSINE TRANSFORM

The DCT is similar to the discrete Fourier transform (DFT) but differs in that it is real, whereas the DFT is complex. Thus the DCT returns a single coefficient for each frequency with a fixed phase, whereas the DFT returns two coefficients (amplitude and phase) for each frequency. The fixed phase of the DCT allows the polarity of each coefficient to be estimated, differentiating between, for example, sloping up and sloping down. We are using the type-II DCT:

$$y(k) = \omega(k)\sum_{n=0}^{N-1} x(n)\cos\frac{k(2n+1)\pi}{2N} \quad k = 0,1,2 \tag{1}$$

where $\omega(k) = \begin{cases} \frac{1}{\sqrt{N}} & k = 0 \\ \sqrt{\frac{2}{N}} & 1 \le k \le 2 \end{cases}$

[*] Thanks to Social Science and Humanities Research Council of Canada (SSHRC), the Fonds de recherche sur la société et la culture (FQRSC), and the Center for Interdisciplinary Research in Music Media and Technology (CIRMMT) for funding.

and N is the length of the signal, x, and k is the number of coefficients.

The 0th, 1st, and 2nd DCT coefficients can be used to capture the broad contours of a signal (in this research, the F_0 trace) that relate to mean, slope, and curvature while ignoring fine details, such as vibrato. In order to transform the DCT coefficients so that they are independent of signal length, the 0th coefficient is divided by $N^{1/2}$, the 1st by $N^{3/2}$, and the 2nd by $N^{5/2}$. After scaling by these terms and another signal-independent constant for each coefficient, the 1st coefficient is approximately in units of cents/second, and the 2nd coefficient is approximately in units of cents/second^2. Additionally, a positive 1st coefficient indicates a negative slope, so its sign is reversed so that it describes positive slope.

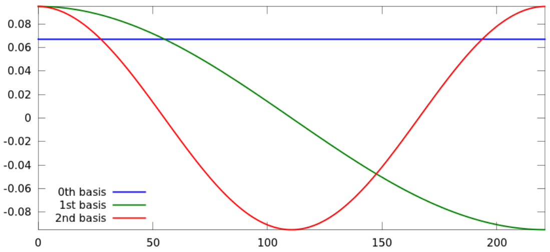

Figure 1: The 0th, 1st, and 2nd DCT basis functions. The blue line represents the 0th coefficient, the green line represents the 1st coefficient, and the red line represents the 2nd coefficient.

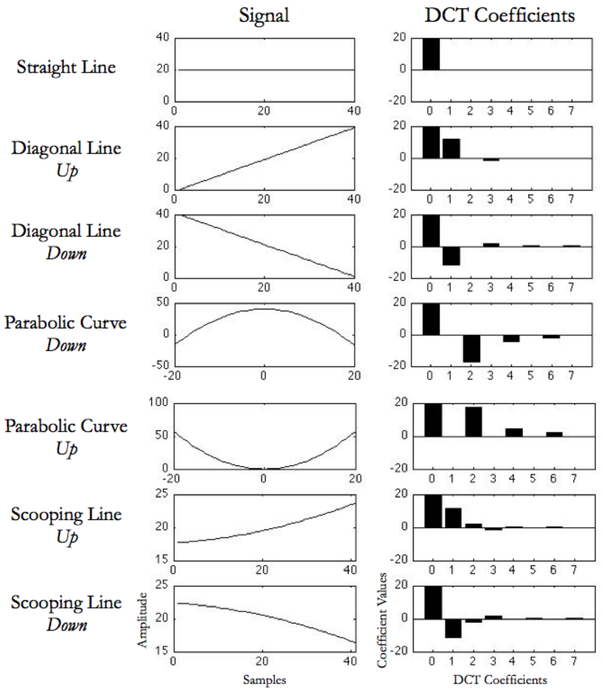

Figure 2: Examples of DCT coefficients (without scaling) for simple signals. The plots on the left are the original signals (straight line, diagonal line, parabolic curve, and scooping line), and the bar graphs on the right are the values for the raw DCT coefficients 0–7, before any transformation as described above.

The 1st DCT coefficient approximates the slope of the evolution of F_0. The slope provides information about whether the singers are gliding up or down or staying relatively stable. The amount and direction of movement (if any) depends on the sign and value of the coefficient. The 2nd DCT coefficient approximates the curvature of the evolution of F_0. The curvature, once the slope has been subtracted, indicates the amount that the F_0 trace is higher or lower in the middle than at the two ends of the time period analyzed. Figure 2 shows the DCT coefficients for seven simple signals: a straight line, diagonal lines up and down, parabolic curves up and down, and scooping lines up and down.

In Figure 2, all of the signals have a mean of 20, so the 0th coefficient remains the same. For the flat line, only the 0th coefficient has a nonzero value since the signal can be completely described by its mean. For the diagonal line, the 1st coefficient has a much larger value than any other (except for the 0th), which demonstrates the relationship between the 1st coefficient and the slope of the signal. For the parabolic curve, the 2nd coefficient has the largest value (again except for the 0th), demonstrating that the 2nd coefficient can be taken to approximate the curvature of the signal. This relationship is more approximate than between the 1st coefficient and the slope, which can be seen in the greater values of the 4th and 6th DCT coefficients for the parabolic curve. The spread of energy across different coefficients occurs because the signals are generated from second-order polynomials rather than cosines.

3. FUNDAMENTAL FREQUENCY TRAJECTORIES

This method assumes that the notes of interest have already been segmented within a recording, using a method such as the one described in [4], and that accurate frame-wise F0 estimates are possible, using a method such as the one described in [5]. As noted above, F_0 trajectories in sung notes can be characterized by variability at the start and end of the note and a stable vibrato section in the middle. The variability in the starting and ending sections of the notes possess some characteristics, described in [6]. Namely the instability at the start of the note often includes an overshoot of the target note while the instability at the end of the note includes some fine fluctuations in the regularity of the vibrato as the note draws to a close in combination with a movement in the opposite direction of the next note to prepare for its arrival. These characteristics of starting and ending sections of the notes may unduly influence the DCT calculations, resulting in the slope and curvature calculations that might not be related to the F_0 trajectory of in the stable vibrato section of the note. In order to prevent this, it is necessary to isolate the stable vibrato section of the note.

In order to evaluate the effectiveness of the DCT for characterizing slope and curvature in F_0 trajectories in sung notes we evaluated the DCT on a set of synthetic signals. Synthetic signals were used instead of F_0 traces from sung notes because they provided ground truth against which to compare the calculated DCT coefficients. The synthetic signals were generated using an Attack-Sustain-Release model. Following from [6], the Attack portion contains an overshoot of the target note, the Sustain portion contains vibrato, and the Decay portion contains a movement in the opposite direction of the next note. Low-pass noise is added to all of the sections in order to simulate some of the randomness found in human performance. The overshoot, preparation, vibrato rate, vibrato depth, and noise standard deviation are all controlled by a series of control points. An example of a synthetic signal is shown in Figure 3.

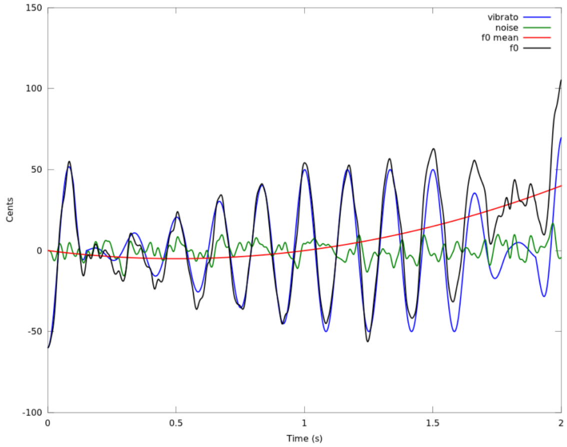

Figure 3: Example of a synthetic F_0 trajectory used in the experiment in this paper. The black line shows the F_0 trajectory composed of the sum of the mean trajectory (red line), vibrato (light blue line), and low-pass noise (green line).

4. APPLYING THE DCT TO F_0 TRAJECTORIES

Before the DCT can be calculated on an F_0 trajectory, the preprocessing step of isolating the stable vibrato section of the note must be performed. We do this using a technique similar to the Empirical Mode Decomposition [7], but designed specifically for the task of eliminating vibrato. All of the peaks and troughs in the F_0 trajectory are identified as the maxima and minima in a 100 ms window that is slid across the signal. The values within 50 ms of these extrema are averaged together to achieve a more robust estimate of the peak values. The midpoints between each pair of extrema are then identified by finding the point in that interval with an F_0 closest to the average of the robustly estimated pitches of the extrema. The stable section of the vibrato is defined as the longest section of the F_0 trajectory where the sequence of extrema are at least 20 cents from their corresponding midpoint, a conservative estimate based on the literature [8]. The starting point of the DCT is the midpoint after the first trough in the stable vibrato section and the ending point is the midpoint before the last peak. Figure 4 shows the results of this analysis on the synthetic F_0 trajectory in Figure 3.

Locating the midpoint between the first and last trough/peak pair in the stable vibrato section of the note is important because the shape of the vibrato can influence the 1st and 2nd DCT coefficient values. This is demonstrated in Figure 5, which shows the results of the DCT calculations from a stable starting point (the start of the stable vibrato section of the note as marked with triangles in Figure 4) to a variable ending point (every sample between the final trough and peak in the stable vibrato section). Both the 1st and 2nd DCT coefficients show a clear influence of the ending position, with both sets of values mirroring the influence of the upwards trajectory of the F_0 values between the trough and the peak. In our algorithm, the F_0 trajectory is smoothed before the DCT is calculated on it in order to minimize the influence of starting/ending point effects and vibrato. The smoothed signal is generated by linearly connecting the midpoints between each trough/peak pair in the stable vibrato section of the F_0 trajectory. This smoothing is especially necessary for shorter notes with only a few vibrato cycles. When only a few vibrato cycles are present, the shape of the vibrato overwhelms the basic shape of the F_0 trajectory.

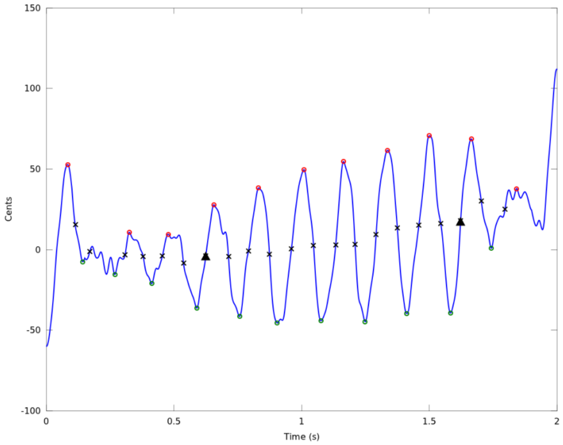

Figure 4: Example algorithm to isolate the stable vibrato section of a synthetic F_0 trajectory. The blue line shows the F_0 trajectory, the red and green circles show the estimated peaks and troughs in the trajectory, respectively, the x's show the midpoint between the corresponding peaks and troughs, and the triangles show the start and end of the stable vibrato section.

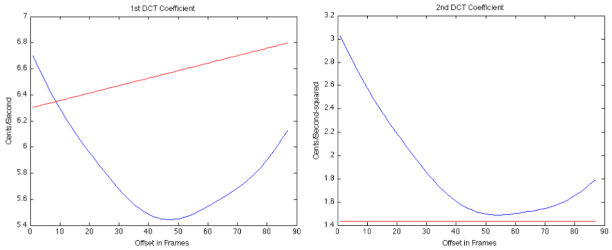

Figure 5: Demonstration of the influence of vibrato on the DCT calculations. Ground truth (red) and estimated (blue) DCT values for a note as the end-point of the calculation is swept over half a vibrato cycle. The plot on the left shows the 1st DCT coefficient and the plot on the right shows the 2nd DCT coefficient.

In order to test the usefulness of the DCT for measuring slope and curvature of various types of notes, we generated 945 synthetic F_0 trajectories. The slope and curvature in these signals fall into seven categories that correspond to those visualized in Figure 2: straight lines, diagonal lines going up and down, parabolic shapes going up and down, and scooping lines going up and down. The overall envelope of the notes fell into three different types, which are characterized by the F_0 activity at the beginning and ending of the notes: flat notes, notes that begin with an upward overshoot of the target note and end with a downward movement, and notes that begin with a downward overshoot of the target note and end with an upward movement. Within these categories and types, three other types of variations were applied. The first was 3 combinations of vibrato rate and depth, which ranged from 35–55 cents for the depths and 5–7 Hz for the rate. The second was 3 different note lengths (1000, 2000, and 3000 ms.). The third was 5 random draws of low-pass noise that was added to the F_0 trajectory. Ground truth for these signals was obtained by calculating the DCT on the unaltered F_0 trajectory used as the basis of the synthetic signals for each of the seven categories (shown in Figure 2). Within these categories the difference between the ground truth and DCT calculations from the full synthetic F_0 trajectories were combined under the three types of note envelopes defined above. Table 1 shows the means and standard deviations of the distance between the ground truth and the DCT calculations for slope (1st DCT coefficient). Table 2 shows the same information for curvature (2nd DCT coefficient).

	GT	**Type 1 (315)**		**Type 2 (315)**		**Type 3 (315)**	
		Mean	*SD*	Mean	*SD*	Mean	*SD*
SL	0	0.8	*0.9*	0.8	*1.1*	0.7	*0.9*
DU	10	0.9	*1.5*	0.8	*0.9*	0.9	*1.1*
DD	-10	0.6	*0.8*	0.8	*0.9*	0.8	*1.0*
PU	-1.0	0.7	*0.8*	0.8	*0.9*	0.9	*1.3*
PD	1.1	0.8	*1.1*	0.6	*0.7*	0.7	*1.0*
SU	7.6	0.8	*1.2*	0.8	*0.9*	0.7	*0.9*
SD	-7.5	0.9	*1.3*	0.9	*1.0*	0.8	*0.9*

Table 1: Means and standard deviations, in cents/second, of the absolute distance between the ground truth and the 1st DCT coefficient, approximating slope, calculated on the synthetic F_0 trajectories. The Types indicated in the first row refer to the three types of F_0 activity at the beginning and ending of the notes. The first column on the left indicates the shape of the F_0 trajectory: straight lines (SL), diagonal lines going up and down (DU, DD), parabolic shapes going up and down (PU, PD), and scooping lines going up and down (SU, SD).

	GT	**Type 1 (315)**		**Type 2 (315)**		**Type 3 (315)**	
		Mean	*SD*	Mean	*SD*	Mean	*SD*
SL	0	1.1	*1.5*	1.6	*3.3*	1.2	*2.1*
DU	0	1.2	*2.0*	1.8	*3.4*	1.6	*3.2*
DD	0	1.0	*1.4*	1.1	*1.8*	1.2	*1.8*
PU	-2.6	1.3	*1.9*	1.2	*2.4*	1.6	*2.8*
PD	2.6	1.7	*3.3*	0.9	*1.3*	0.9	*1.3*
CU	1.3	1.6	*2.5*	1.2	*1.7*	1.0	*1.8*
CD	-1.3	1.5	*3.8*	1.1	*2.1*	1.6	*2.5*

Table 2: Means and standard deviations, in cents/second^2, of the absolute distance between the ground truth and the 2nd DCT coefficient, approximating curvature, calculated on the synthetic F_0 trajectories. The Types indicated in the first row are the same as in Table 1.

The results in Tables 1 and 2 show that the DCT is robust for these types of signals across a range of note shapes and envelopes. The means and standard deviations are all around 1 unit, which indicates that while there is some influence for the increased complexity of the synthetic signal over the simple lines shown in Figure 2, the impact is minimal. The 1st DCT coefficient calculations prove to be more robust than the 2nd DCT coefficient calculations, since the means and standard deviations of the absolute differences between the calculations and the ground truth are much smaller than the ground truth values for the 1st DCT coefficient but closer in size for the 2nd.

5. APPLICATIONS

The use of the DCT to characterize F_0 trajectories has applications in two main areas. The first is in the study of singing performance practices. The DCT coefficients provide useful generalized descriptors that allows for comparisons not only across different notes in the same performance, but also across notes in different performances by the same singer and across different singers. The second is in the area of singing synthesis. While much work has been done on the timbre of the singing voice and, more recently, the pitch and amplitude variations related to vibrato, the issue of F_0 evolution over the duration of a note has received substantially less attention. The DCT coefficients are a useful supplement to mean pitch when specifying tuning-related information in digital re-creations of the signing voice.

6. CONCLUSIONS

This paper has shown that the use of the DCT on F_0 trajectories provides a good summary of pitch-related characteristics by measuring its slope and curvature. It also presents an algorithm for smoothing F_0 trajectories in order to address potential issues with vibrato and start/endpoint effects. The DCT was evaluated on smoothed versions of 945 synthetic F0 trajectories, generated from a model of the singing voice. The results of this evaluation showed that the DCT is a good way of measuring slope and curvature, with the slope calculations being more robust than the curvature calculations.

7. ACKNOWLEDGMENT

The authors would like to acknowledge the assistance of Dan Ellis in the preparation of this paper.

8. REFERENCES

[1] J. Sundberg, "The perception of singing," in *The Psychology of Music*, D. Deutsch, Ed., 2nd ed San Diego, CA: Academic Press, 1999, pp. 171–214.

[2] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[3] T. Baran*, et al.*, "Preserving the character of perturbations in scaled pitch contours," in *ICASSP*, 2010, pp. 417–420.

[4] J. Devaney*, et al.*, "Improving MIDI-audio alignment with acoustic features.," in *WASPAA*, 2009, pp. 45–48.

[5] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *JASA,* vol. 111, pp. 1917–30, 2002.

[6] T. Saitou, *et al.*, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," *Sp. Comm.,* vol. 46, pp. 405–17, 2005.

[7] N. E. Huang, *et al.*, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. London. Series A.,* vol. 454, pp. 903–995, 1998.

[8] E. Prame, "Vibrato extent and intonation in professional western lyric singing," *JASA,* vol. 102, pp. 616–21, 1997.