Proceedings of Meetings on Acoustics

Volume XX http://acousticalsociety.org/

169th Meeting of the Acoustical Society of America

Pittsburgh, Pennsylvania

20 May 2015

Technical Area: Paper XXXXX

Using spectral analysis to evaluate flute tone quality

Ron Yorita - California Polytechnic State University
Computer Science Department, San Luis Obispo, CA; ron.yorita@gmail.com

Dr. John Clements - California Polytechnic State University
Computer Science Department, San Luis Obispo, CA; clements@brinckerhoff.org

Many skilled flutists place a high priority on good tone quality, or timbre. Unlike pitch and rhythm, timbre is difficult to objectively quantify. This project explores (1) how tone quality is described by skilled flutists, (2) whether the harmonic spectrum has some correlation with tone quality, (3) whether certain harmonic spectra are preferred, or considered "good."

Thirty-one flutists ranging from high school students to professionals were recorded. A set of samples was used in surveys and interviews to capture descriptors and ratings of tone quality. All of the recorded samples were analyzed using application programs, Harmonic Analysis Tools (HAT), created for this study. HAT uses digital signal processing techniques to produce "spectral signatures." The signatures consist of the harmonic content, pitch, and amplitude of a sample.

The outcome of this research is a baseline set of some frequently-used descriptors. In addition, results showed some correlation between harmonic spectra and descriptors. There were also trends in preferences with respect to certain spectral characteristics. An unexpected finding was that University students showed divergent timbre preferences compared to highly experienced flutists.

1 Introduction

For the purposes of this paper, the terms "tone quality," "tone color," and "timbre" will be used interchangeably. These terms will be defined as: the audible difference in character that a listener perceives for two notes played at the same pitch [**?**, **?**]. For example, an oboe playing the note at a given pitch has a different timbre than a flute playing the same pitch. Moreover, two different flutists playing the same pitch can have different timbres or tone qualities.

Flutists value good tone quality, and some regularly invest time practicing tone exercises. Geoffrey Gilbert suggested spending one-third of practice time on tone studies [**?**]. Some of Gilbert's former students include James Galway, William Bennett, and Trevor Wye. There is a substantial corpus of information devoted to flute tone quality. These sources include: books, periodicals, websites, fora, and scholarly papers. Browsing the bibliographies of these, dissertations, or quality websites on the topic of flute timbre reveal a wealth of literature [**?**, **?**, **?**].

When flutists talk about tone quality, there is often an implicit understanding that different musical contexts require particular tone colors [**?**, **?**, **?**, **?**]. For example, what is appropriate for an early Baroque chamber music piece might not be appropriate for an Ian Anderson (Jethro Tull) improvised solo. Further, within a given genre or even within a single composition, circumstances may call for a variety of subtleties of timbres. Appropriate tonal variations are applied to enhance musical interpretation during a performance. This might be akin to a painter using a palette with a variety of subtle hues to shade his art.

In addition to musical circumstances, personal taste plays a role in evaluating tone quality. Some people prefer Jean-Pierre Rampal, some James Galway, and others Emmanuel Pahud. Interestingly, there have been studies that show personality types influence timbre preferences [**?**, **?**]. Another aspect of preference may be the level of ear-training and the aural acuity of the listener. Cultural background is another factor that influences tone perception [**?**].

Although musical context, personal taste, and other factors influence each listener's assessment of "good" flute tone quality, there may be some consensus on standards for tone. For example, a beginning elementary school flutist will probably produce a sound that is less desirable than the principal flutist from an elite symphony orchestra.

Flute students may find it challenging to develop their sound while navigating through the aforementioned subjectivity. This is further exacerbated by perplexing descriptions of tone color: *bright*, *dark*, *dull*, *edgy*, *hollow*, *round*, *fuzzy*, *pure*, *reedy*, etc.[1] It is also not uncommon to find flute literature indicating that more harmonics enriches tone quality. However, the specifics of which harmonics and the appropriate balance is seldom specified. Even when data like harmonic spectra are used to illustrate differences in tone quality, readers must often rely on only written descriptions of timbre differences without the benefit of aural input [**?**]. To quote Roger Stevens: "Verbal terms describing tone colors are quite inadequate, and as such descriptions are, for the most part, purely subjective." [**?**]

This project explores (1) how tone quality is described by skilled flutists, (2) whether the harmonic content or spectral signature has some correlation with tone quality, (3) whether certain acoustic signatures are preferred, or considered "good." The hope is that there are some measurable aspects of timbre that can be associated with desirable qualities.

In order to accomplish these objectives, the first step was to obtain a range of flute tone samples. Next, these samples were analyzed by application programs written specifically for this project. Finally, descriptors and ratings from experienced flutists were procured via online surveys and one-on-one interviews.

In section 2, we briefly discuss background, both for flute construction and spectral analysis. Section

[1] Here and for the remainder of the paper, terms used to describe flute tone timbres will be italicized.

3 summarizes related work. In Section 4, we discuss the methodology, both for the collection of recorded flute tones and for the following survey. The results of the survey appear in section 5. Finally, we conclude in section 6.

2 Background

Flute tones are produced when the air column inside the flute vibrates in a uniform manner, propagating outward as sound waves. Acoustically, the flute acts as a pipe open at both ends (whose length is determined by the configuration of the keys), meaning that the vibration produces a "fundamental" frequency whose wavelength is twice the length of the tube [**?**]. In addition to this fundamental, though, integer multiples of this frequency are also produced. For example, many people know that the note A440 vibrates at 440 cycles per second, or 440 Hz. The fundamental frequency, also known as harmonic 1 (H1), for this note is 440 Hz. There is a harmonic that vibrates at 440 Hz x 2 = 880 Hz, or the second harmonic (H2). Harmonic 3 (H3) has frequency 440 Hz x 3 = 1,320 Hz, and so forth. The combination and balance of these harmonics determine the tone quality, or "timbre" [**?**, **?**, **?**, **?**, **?**].

There is some debate into how much the flute construction material influences flute timbre. Scientific experiments have demonstrated that different the materials (e.g., silver, gold, platinum, wood, copper, cardboard, concrete) have no effect on perceived tone quality [**?**, **?**, **?**, **?**]. However there are proponents flutists that advocate material does make a difference [**?**, **?**]. Regardless of this difference of opinion, there is consensus that the flutist has a large influence on tone quality [**?**, **?**, **?**, **?**].

The flutist and the flute are one factor of tone quality. Facilitating analysis for this study required recording notes from different musicians. Practicalities necessitated recording subjects at different venues [**?**]. A rough rule of thumb for factors influencing recordings is: 50% musician (and flute), 20% room, 20% microphone position, 10% microphone choice [**?**]. The factors that could be controlled during recording sessions were. Microphone placement was always approximately 5 feet directly in front of the musician. The identical microphone and recording equipment were used for all the sessions. The recording level was identical for all sessions.

Perhaps surprisingly, the variability of recording environments between flutists is not a major issue for this study. To see this, consider a recording session in which microphone position caused attenuation of a certain band of frequencies, and suppose further that experts were found to dislike these recordings. For our purposes, this represents valuable information! That is, we are trying to discover what *recordings* are found to be pleasing, and not to attribute these differences specifically to the flutist, the flute, or the recording session. The essential relationship is that the recording used to produce the spectral signature is what a listener uses to describe the quality.

One caveat remains; we did not have precise control of the reproduction setup used by survey participants. If one population consistently used, for instance, headphones with poor reproduction of frequencies between 1 and 2 kHz, this might affect our results.

Application programs, or Harmonic Analysis Tools (HAT), were implemented to analyze the recorded flute tones and produce spectral signatures [**?**]. HAT uses Standard Digital Signal Processing (DSP) techniques. For example, spectra were produced using Fast Fourier Transforms, and pitch was discovered using autocorrelation.

The HAT applications can be used to analyze spectra for any instrument, but were tailored to the flute. Flute tone is characterized as relatively pure, with few and weak upper harmonics [**?**]. Some of the studies cited in the Related Work section showed that the harmonics above H7 are generally very weak, or nearly absent. Observations using Real Time Analyzers (RTAs) confirmed this. Therefore, most of the HAT display

modes only render the first seven harmonics. Figure 1 shows the colors used to represent each harmonic. The convention is that the fundamental (H1) uses a white line. Octaves of the fundamental (H2 and H4) are yellow. The harmonic that is an octave and a fifth higher than the fundamental (H3) or two octaves and a fifth higher than the fundamental are green (H6). Thirds are red (H5), and the minor seventh (H7) is purple.[2]

Harmonic #	Note or pitch	Graph color
H1	Fundamental (perceived pitch)	White
H2	Octave above fundamental	Yellow
H3	Octave + fifth above fundamental	Green
H4	Two octaves above fundamental	Yellow
H5	Two octaves + third above fundamental	Red
H6	Two octaves + fifth	Green
H7	Two octaves + minor seventh above fundamental	Purple

Figure 1: HAT Harmonics Color Coding

Figure 2 shows a HAT spectral signature. The top right portion of the screen displays the target pitch that subject "C10" was asked to play (g4 or 392 Hz) along with the actual average pitch measured by HAT. The left half of the screen shows the harmonics over time. The vertical axis indicates how strong each harmonic is. In this example, the flutist's H1 (white graph line) is very strong and H2 (yellow) is also present. The vertical lines indicate one-second intervals so the overall duration of this note was approximately 3.5 seconds. This note was played without vibrato, what is called "straight tone." Fluctuations in the harmonics are evident. The right side of the screen shows pitch (top-right graph) and loudness (lower-right graph). Both of these graphs also have vertical reference lines at one second intervals.

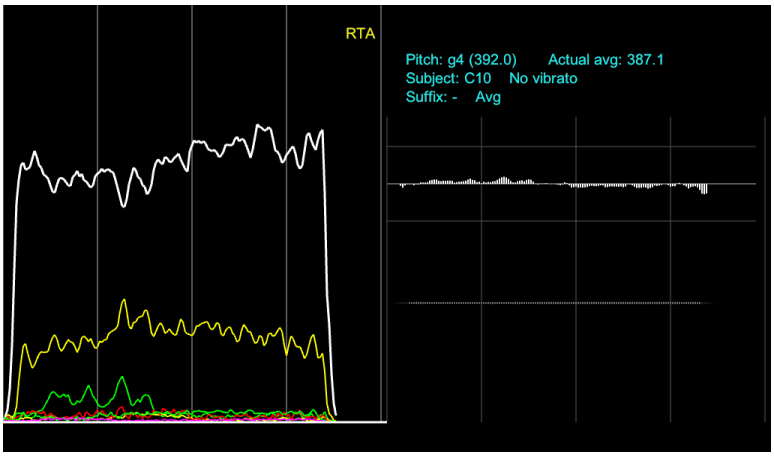

Figure 2: HAT Spectral Signature

[2]For simplicity of interpretation, we elected to use the same color to highlight harmonics that are separated by octaves in this display. The full tool provides another display mode that separates these harmonics.

3 Related Work

There is a body scientific study that has used spectral analysis of flute tone quality. Dr. John Coltman [**?**] investigated Just Noticeable Differences (JND) in perceived tone quality by incrementally varying harmonics. Coltman also published a study analyzing how a flutist's airstream can determines harmonic content [**?**]. Dr. Neville Fletcher demonstrated that the fundamental may be weaker than upper harmonics on lower flute range [**?**]. In that study, he also showed that the balance of harmonics changes for different pitches. In another study, Fletcher analyzed vibrato, and noted that the mix of harmonics can vary at different phases of vibrato [**?**].

There are also studies that have used spectral analysis to help musicians improve tone quality. Dr. Billington analyzed how the flutist's physical configuration can affect timbre [**?**]. He posits how different mixtures of harmonics influence *brightness* and *fullness* in tone color. Dr. Rundus examined various factors influencing vocal quality and how to apply the information provided by the spectral analysis including: onsets and releases, breathing for singing, resonance, focus of the tone, articulation, and musical expression [**?**]. Daniel Jones conducted research with middle school trumpet players [**?**]. By using a real-time spectrogram for visual feedback he observed an improvement in 8th grade students' tone quality. The paper cites several other studies using spectrograms to facilitate tone quality improvement.

4 Methodology

For our study, thirty-one flutists of various skill levels were recorded. The flutists included: four high school students, 16 university students (both music majors and non-majors), and 11 professionals. Each participant read and signed an informed consent form prior to being recorded. They were instructed to play a series of long-tones in a prescribed manner. We captured and analyzed 1,600 notes.

A set of twelve straight-tone samples grouped into pairs (with the same pitch) was selected for use in an online survey. The selection set is small for practical reasons. Preliminary studies revealed that listener fatigue degraded feedback quality when comparing large numbers of samples. To minimize the impact of listener fatigue, the number of comparisons needed to be limited. Restricting the comparisons to sixxo pairs allowed evaluators to complete the survey process in approximately 10-15 minutes. Keeping the survey short ensured a high completion rate with attentive responses throughout. The obvious trade-off is that data could only be gathered for a small set of tone samples.

The twelve samples were selected based on their acoustic signatures. An important criterion was to use samples with relatively stable harmonics over the duration of the note. The other selection criterion was visual and aural distinctiveness within a given pair of notes. Following these criteria, tone-pairs had:

- visually different harmonic signatures
- timbre differences that most skilled listeners could easily hear
- the same pitch or note name

There were two factors behind our decision to use only straight-tone notes. The first is that vibrato adds significant complexity to the harmonic characteristics of a long-tone. Some observations about vibrato and harmonics are discussed in the original study write-up [**?**]. The second factor is that tone samples with vibrato might draw evaluators to focus on the vibrato quality rather than the underlying tone quality.

The survey was intentionally designed to be somewhat open-ended. Specifically, participants were asked to listen to, and then describe/rate tone quality for notes without any guidance. The musical context, as well as any other criteria for judging timbre were omitted from the instructions. One of the objectives was

to observe whether there is some level of consensus about flute tone quality independent of context. No examples of adjectives or phrases for tone quality were provided. This avoided biasing evaluators, and required them to use their own descriptors. The instructions encouraged participants to freely express their opinions as this is subjective and therefore there are no correct or incorrect answers. In order to access the survey, all respondents were required to read and acknowledge an informed consent.

Two sets of survey results are presented in the following section. The first set, and the primary target demographic, consists of experienced flutists. The second set consists of University students that have had some musical training.

The primary group are members of the Flute List. This is the "longest-established internet mailing list relating to the flute" [**?**]. There were 121 Flute List members participating in the survey, and the results from a subset consisting of 41 individuals were selected. These 41 individuals had: 10 or more years of teaching experience, and 10 or more years of private lessons, and play/practice/rehearse 10 or more hours per week. Essentially, they are seasoned instructors with substantial private training that actively maintain their performance skills. This group will be referred to as the FL10s (Flute List 10's).

The second group consists of University students from California Polytechnic State University, San Luis Obispo, California. Of the 131 students that participated in the survey, 66 students indicated that they had had musical training. These 66 musicians will be referred to as the CPM (Cal Poly Musicians).

5 Results

5.1 Descriptors

The nouns "descriptors," "adjectives," and "terms" will be used interchangeably here. A set of frequently used descriptors was extracted from the FL10s survey responses to understand how "skilled flutists" describe tone quality. These terms were then categorized with their ratings to determine if they are considered favorable or unfavorable qualities. Any patterns revealed here apply to the survey samples, and may not be generally applicable.

The ten most frequently used descriptors are shown in table 1. The rating scale range was from 1 ("poor") to 5 ("great"). A 3 would be considered a *neutral* rating. *Favorable* descriptors accompanied ratings of 4 or 5; *Neutral or favorable* descriptors were used with ratings of 3 or higher; *Unfavorable* descriptors were used with ratings of 2 or lower. The *Across all ratings* category is for descriptors accompanying all ratings (rating from 1 to 5). The entries are ordered from most to least frequently used.

Columns 1 and 2 are self-explanatory. Column 3, *Count*, shows the number of times each descriptor was used by the FL10s. Column 4, *#People*, indicates the number of distinct FL10s participants who used that descriptor. Together, column 2 and 4 provide some indication of whether some individuals repeatedly used a particular descriptor across the survey samples. The last column gives a sense of whether the descriptor might be considered positive, negative, or non-determinant.

Figure 3 provides some additional context by showing the FL10s rating distribution. There were somewhat more unfavorable ratings resulting in an average of 2.76. The "great," or 5, was given sparingly relative to the other ratings.

Another way to organize tone descriptors is based upon ratings. Table 2 shows sets of adjectives for various rating categories. The parenthetical numbers indicate the number of times each term was used across all the survey samples. This table contains 34 descriptors, and all of them were used at least five times.

Although the terms *bright* and *dark* seem to have opposite meanings, they appear to be used inter-

Table 1: FL10s Descriptor Usage

Ranking	Descriptor	Count	#People	Category
1	focused	52	22	neutral/favorable
2	airy	37	17	neutral/unfavorable
3	unfocused	31	19	unfavorable
4	edgy	29	16	across all ratings
5	clear	22	12	neutral/favorable
6	rich	20	12	favorable
7	weak	20	9	unfavorable
8	full	18	14	favorable
9	diffuse	15	9	across all ratings
10	open	14	8	across all ratings

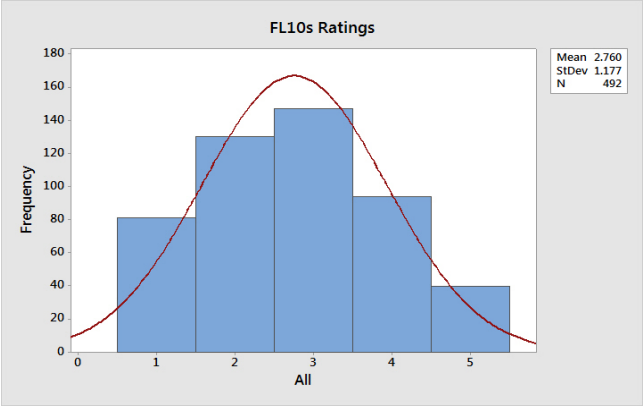

Figure 3: FL10s rating distribution

changeably describing samples rich in harmonics. Terms like *sharp* (pitch vs. point/edge) and the survey comments did not always provide sufficient context to clearly interpret what was intended. The *Across all ratings* category may reflect personal taste. For example, some individuals may or may not prefer an *edgy* or *diffuse* tone quality.

Another way to view the descriptors is to focus on the terms associated with either "great" or "poor" ratings. Since these ratings are at the extreme ends of the scale, they indicate either very positive or very negative reaction to a timbre. For the FL10s, the "great" rating occurred 40 times, and the top three descriptors were *focused*, *clear*, and *rich*. The "poor" rating occurred 81 times, and the most frequently used descriptors were *airy*, *unfocused*, and *weak*. Table 3 shows the frequency of these terms. The percentage indicates how often the term was used. For example, there were 22 occurrences of the term *focused* within the 40 "great" ratings; $22 \div 40 = 55\%$.

Within the 12 survey samples, an important criteria for the FL10s appears to be whether the timbre sounded *focused* or *unfocused*. Samples perceived as *focused*, *clear*, or *rich* were rated positively. Samples that were *airy*, *unfocused*, or *weak* were rated negatively.

5.2 FL10s Ratings

The average rating for each sample provides a high level view of tone quality preference. These are shown in figure 4. Both the complete Flute List and the FL10s averages are included, and the results are similar. Some of the ratings, like 1B, 2A, 3B, and 4A, are slightly lower for the FL10s. A few of the others are slightly higher.

The FL10s results ranged from 1.93 to 4.07. If 3 is considered average, 8 of the 12 samples were below

Table 2: FL10s Descriptors

Circumstances	Descriptors
Favorable	rich(20), full(18), resonant(5), colorful(4)
Neutral or favorable	focused(52), clear(22), round(8), bright(8), dark(7), buzz(5)
Neutral or unfavorable	airy(37), forced(14), harsh(13), hollow(12), soft(12), dull(12), overblown(10), brassy/trumpety(6), lacking-core(6)
Unfavorable	unfocused(31), weak(20), thin(14), unsupported(11), sharp(9), breathy(8), muffled(6), nasal(6), uncontrolled(6), uncentered(5)
Across all ratings	edgy(29), diffuse(15), open(14), loud(11), warm(5)

Table 3: FL10s Descriptors for "great" or "poor" ratings

Descriptor	Occurrence	Percent
focused	22	55%
clear	8	20%
rich	8	20%

Descriptor	Occurrence	Percent
airy	13	16%
unfocused	12	15%
weak	10	12%

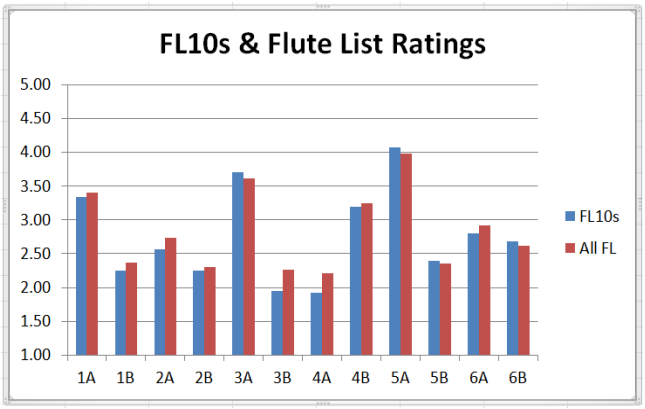

Figure 4: FF10s and Flute List ratings

average, and the remaining 4 were above average.

The FL10s ratings are summarized in table 4. The results are sorted by lowest to highest rating. There are two categories of timbre that FL10s rated unfavorably:

1. Tones perceived as *weak*, *airy*, or *unfocused* (4A, 3B, 1B, 5B, 2A, 6A).
2. Tones perceived as excessive in some form: *overblown*, *edgy*, or *brassy/trumpety* (2B, 6B).

Sample 4B had a range of responses and was rated favorably. Although it was sometimes described as forced or trumpety, it also received favorable remarks like rich or focused. The remaining three samples (1A, 3A, 5A), were rated favorably and had descriptors like: focused, dark/bright, rich, or clear.

The following subsections examine the spectra, ratings, and descriptors in greater detail. The first subsection looks at the pitch G4 which has the two lowest rated samples. Then pitch D4 is analyzed since it has the next two lowest rated samples. The last subsection continues to B4 and D5 in a similar vein.

5.2.1 G4 Spectra

The two lowest rated samples have a target pitch of G4. Figure 5 shows the harmonic signatures for all G4 samples. The images are ordered left to right, top to bottom based on rating. Samples 4A and 3B were rated the lowest and neither received any "great" ratings. They share some common descriptors for the unfavorable ratings (2-ratings plus "poor" ratings): *unsupported*, *weak*, and *unfocused*. They were never

Table 4: FL10s summary of ratings and descriptors

Rating	Sample	Pitch	Great	Poor	Descriptors
1.93	4A	G4	0	16	airy, weak, uncontrolled
1.95	3B	G4	0	14	unfocused, breathy/airy, thin
2.24	1B	D4	1	12	slightly-diffuse vs. unfocused, weak, airy
2.24	2B	D4	0	11	sharp, overblown, edgy
2.39	5B	B4	2	6	unfocused, weak, airy
2.56	2A	D4	2	8	unsupported, airy, thin
2.68	6B	D5	1	8	refined vs. nasal, brassy/trumpety, edgy
2.80	6A	D5	1	3	bright vs. unfocused, weak, airy
3.20	4B	G4	4	3	rich, focused, full vs. forced, nasal, trumpety
3.34	1A	D4	4	0	focused, dark/bright, rich
3.71	3A	G4	10	0	focused, clear, rich
4.07	5A	B4	15	0	focused, clear, full

described as *dark* or *bright*. Visually, both have a strong H1 (white). Sample 3B also has a very strong H2 (yellow). Both have very little H3 (green) or H5 (red). There is a gap between the strongest harmonics and the upper harmonics.

Figure 5: FL10s g4 spectra ratings

In contrast, 3A was the second most favored sample in the survey. The harmonic signature is very rich in harmonics. This is a case where the H2 and H3 are stronger than the fundamental. Some of the "great" descriptors were: *focused*, *clear*, and *rich*. Some of the other favorable terms include *full*, *round*, and *bright*/*dark*.

Interestingly, 4B has diverse results. Individuals that liked 4B used terms like: *rich*, *focused*, and *full*. Those that dislike 4B use adjectives like: *forced*, *nasal*, and *trumpety*. Another descriptor often used for

4B was *edgy*. H1 and H3 are very prominent in the harmonic signature, and all of the harmonics fluctuate considerably. The room used for recording 4B's sample had considerable reverb. It is not certain if this contributed to the fluctuating harmonics.

5.2.2 D4 Spectra

The third and fourth lowest rated samples have a target pitch of D4. Both 1B and 2A were rated unfavorably and share some common descriptors with the lowest rated G4 samples (*airy*, *unfocused*, *thin*). For 1B and 2A, either H1 or H2 is very strong, and then there is a gap. The upper harmonics comprise a relatively small portion of the mix for these two samples. D4 is near the bottom of the flute range where some flutists' spectra have a rich mix of upper harmonics. Neither of these two samples exhibits this characteristic.

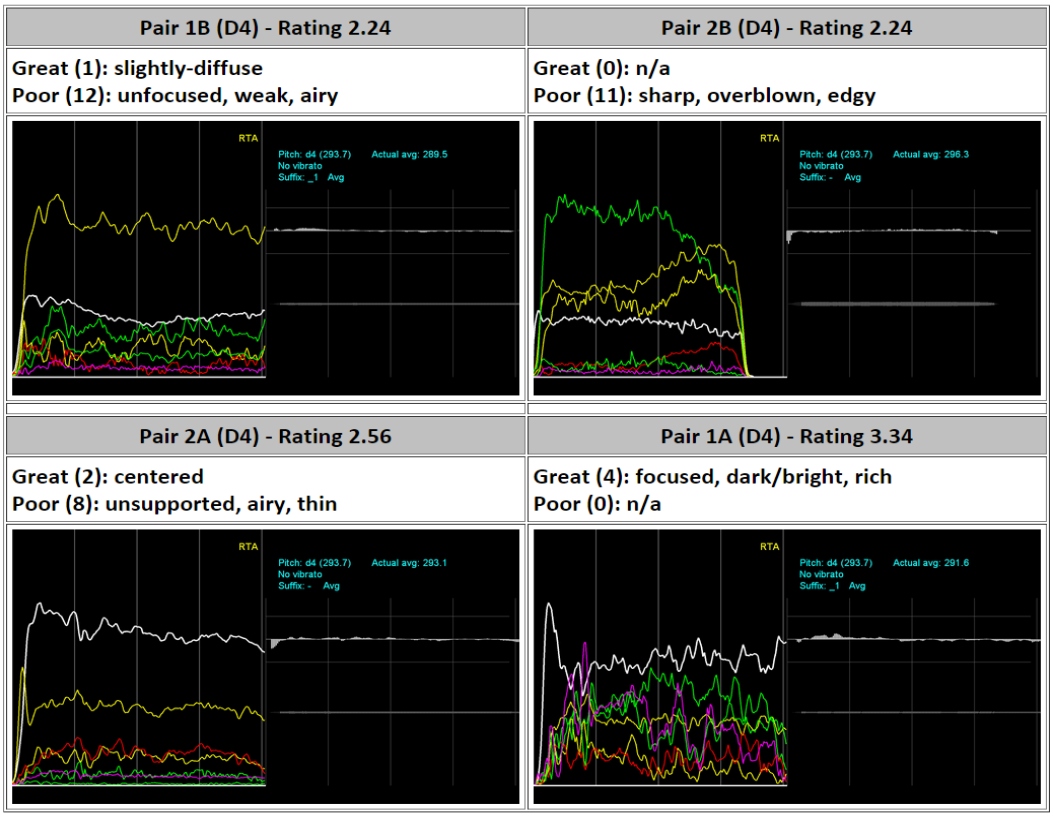

Figure 6: FL10s d4 spectra ratings

Sample 2B was also rated unfavorably. Unlike the samples that were disliked for being *airy* or *unfocused*, 2B was judged as being *sharp*, *overblown*, and *edgy*. There is an unusually strong H3 dominating the signature. In this case, H1, H2, and H3 are all stronger than the fundamental.

Sample 1A was the only D4 sample that was rated somewhat positively. Some of the favorable descriptors included: *focused*, *dark/bright*, and *rich*. The term *edgy* was also used for this sample. Visually, the spectrum is rich in harmonics. H7 (purple) is unusually strong.

5.2.3 B4 and D5 Spectra

The target pitch B4 has one unfavorably rated sample and the other sample was the highest rated in the survey. Sample 5B received mainly unfavorable or neutral ratings. Like some of the other samples with low ratings, 5B has descriptors like: *unfocused*, *weak*, and *airy*. It can be characterized as having a dominant H1 with little presence of upper harmonics.

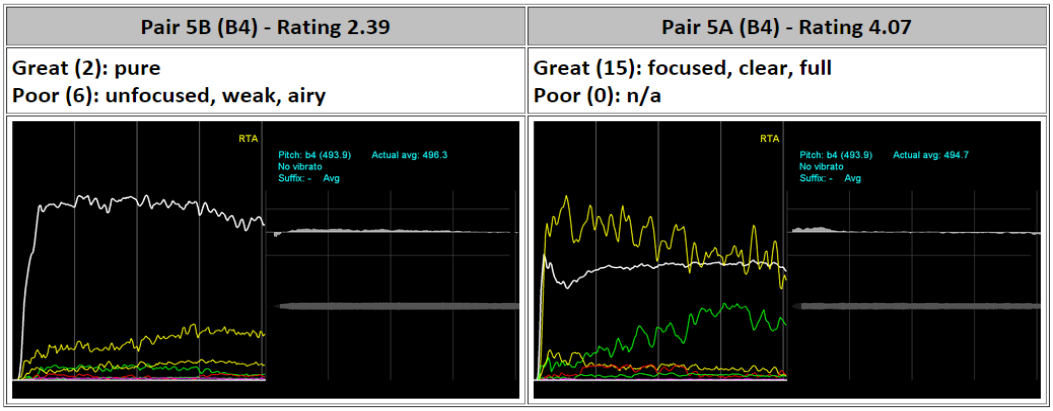

Figure 7: FL10s survey results for pitch B4

Sample 5A had the highest ratings with "great" descriptors like: *focused*, *clear*, and *full*. Other adjectives include: *vibrant*, *rich*, and *resonant*. It is tempting to visually compare the spectrum for 5A with 3B as they share some similarities, but 5A has a greater presence of H3. However, since 3B is a lower pitch it is probably not appropriate to correlate the signatures.

The final pair of notes had D5 as the target pitch. Sample 6B contains an unusually strong H3. Among the 31 flutists, it was very rare to see H3 as the dominant harmonic for pitches in the second flute octave. Although the ratings were generally neutral to unfavorable, more than 25% of the ratings were favorable. Those that liked 6B often used descriptors like: *focused*, and *clear*. The negative terms included: *nasal*, *brassy/trumpety*, and *edgy*. This is a case where some common descriptors accompanied both favorable and unfavorable ratings. The adjectives edgy and *brassy/trumpety* are examples of this type of descriptor.

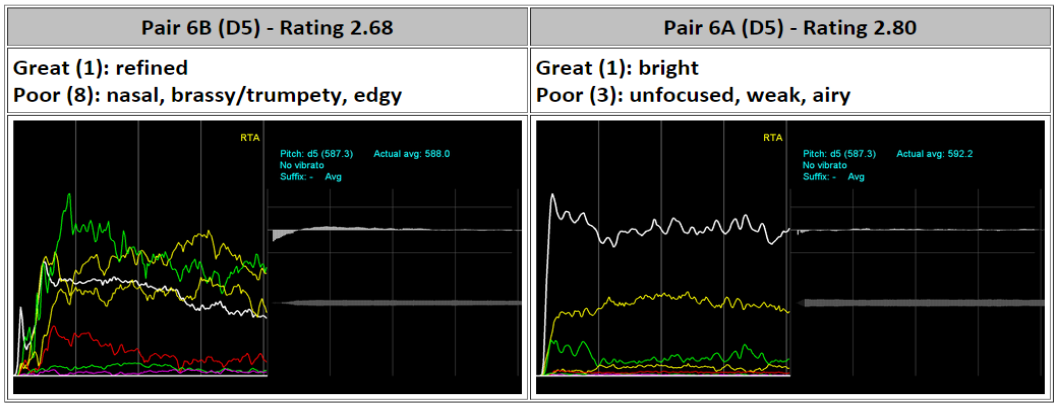

Figure 8: FL10s survey results for pitch D4

Sample 6A was close to neutral in terms of ratings. There were some interesting contradictions in descriptors: *focused* versus *unfocused*, *bright* versus *dull/pale*. The bulk of the adjectives were: *unfocused*,

weak, and *airy*.

5.2.4 Trends in Ratings and Harmonic Signatures

Since there are only 12 samples, and these are subdivided into 5 different target pitches, it is not prudent to make sweeping generalizations from these results. However, there are trends that can be observed within the survey data:

- Samples with strong H1 and/or H2 and relatively weak upper harmonics (H3-H7), were not rated highly by experienced flutists. Visually, there is a gap between the prominent harmonics and the upper harmonics. Samples of this type were seldom described as *bright*/*dark*, or *edgy*. Rather, these samples are often described as *weak*, *airy*, or *unfocused*.
- Samples with a strong H3 were described as *edgy*, *nasal*, or *trumpety*/*brassy*. When H3 was disproportionate, the samples received negative ratings.
- Samples with a balance of harmonics received favorable ratings.
- For the FL10s, the descriptors *bright*/*dark* were used mainly for the top 2 samples. However, the full Flute List used these terms more liberally for other samples containing a high level of H3 (2B, 6B, 4B, 5A, 1B).

5.3 Cal Poly Ratings

Figure 9 compares FL10s and Cal Poly musicians (CPM) ratings. The samples are sorted according to FL10s rating results. There are clearly differences between FL10 and CPM preferences. The overall mean for CPM is 3.17 which is higher than the FL10s mean of 2.76. It is not too surprising that FL10s has a lower average; the FL10s are flute instructors and would be expected to have a critical ear when evaluating flute timbre.

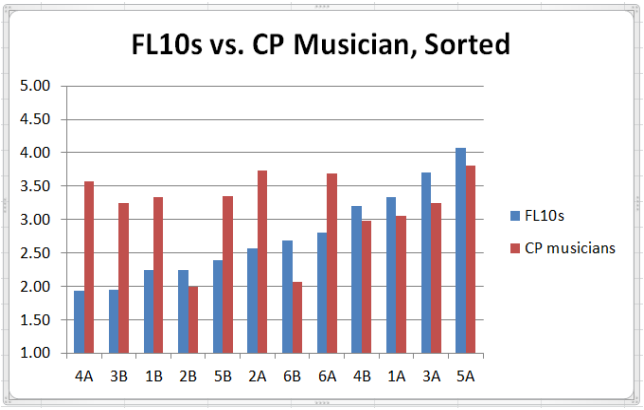

Figure 9: FL10s vs. Calpoly sorted ratings

Surprisingly, all the samples that the FL10s described as *weak*, *airy*, *unfocused* (4A, 3B, 1B, 5B, 2A, 6A) were more acceptable to CPM. In fact, all of these samples received unfavorable rating from FL10s, but received favorable ratings from the CPM. For example, the FL10s lowest rated sample, 4A, received a favorable rating by the CPM. The two-sample t-test was used to calculate the statistical significance of the rating difference between the FL10s and CPM (see table 5). A common rule of thumb when searching for

Table 5: FL10s and CPM ratings and p-values

Sample	Pitch	FL10s	CPM	Difference	p-value
4A	G4	1.93	3.58	-1.65	0.000
3B	G4	1.95	3.24	-1.29	0.000
1B	D4	2.24	3.33	-1.09	0.000
2B	D4	2.24	1.98	0.26	0.173
5B	B4	2.39	3.35	-0.96	0.000
2A	D4	2.56	3.73	-1.17	0.000
6B	D5	2.68	2.06	0.62	0.005
6A	D5	2.80	3.68	-0.88	0.000
4B	G4	3.20	2.98	0.21	0.355
1A	D4	3.34	3.06	0.28	0.137
3A	G4	3.71	3.24	0.46	0.027
5A	B4	4.07	3.80	0.27	0.117

statistically significant results is that we want a false result in fewer than 1 out of 20 cases. When running multiple tests, we must adjust our threshold p-value to ensure that a false positive in *any* measurement occurs with likelihood less than 1 in 20. Since we're performing 12 measurements, this corresponds to a value of $1 - 19/20^{1/12}$, or 0.00427. For this set of samples (4A, 3B, 1B, 5B, 2A, 6A), all of the p-values round to 0.000, clearly indicating statistical significance.

In contrast, the two samples that the FL10s describe as *overblown* or *edgy* (2B, 6B), received lower scores from the CPM. The rating difference for 2B and 6B are not significant. All of the samples that FL10s rated favorably (4B, 1A, 3A, 5A) received slightly lower ratings from the CPM. None of these can be considered statistically significant.

The results suggest that CPM preferred the tones with less harmonic content than FL10s. Within the context of the survey results, highly trained and skilled flutists have a different criteria for tone quality than CPM. This raises some interesting questions:

- Do only highly trained/skilled flutists really appreciate the tone quality of accomplished flutists?
- Does the general public actually prefer flute tones with less harmonic content?

6 Conclusions

A set of frequently used descriptors emerged from the FL10s' survey results. These 41 flutists have significant teaching experience, received considerable private training, and have maintained their playing skills. Within the context of the 12 survey samples, a set of descriptors emerged. The top ten descriptors, ordered by frequency, are:

1. *focused*
2. *airy*
3. *unfocused*
4. *edgy*
5. *clear*
6. *rich*
7. *weak*
8. *full*
9. *diffuse*
10. *open*

There are associations between these descriptors and tone preference. The most commonly used descrip-

tors for favorably rated samples are: *focused*, *clear*, and *rich*. The unfavorably rated samples are generally accompanied with descriptors like: *airy*, *unfocused* and *weak*.

There are some adjectives that were used across all ratings: *edgy*, *diffuse*, *open*, *loud*, and *warm*. These terms may represent attributes that reflect personal taste.

There are associations between spectral characteristics and some descriptors:

- Samples that contain a balance of harmonics were described as: *focused*, *rich*, *bright*/*dark*, or *clear*.
- Samples lacking upper harmonics have descriptors like: *airy*, *unfocused*, *weak*, or *thin*.
- Samples with unusually strong H3 have adjectives like: *edgy*, *nasal*, *brassy*/*trumpety*, or *forced*.

There are two categories for the unfavorable tones:

- Samples with few upper harmonics (H3-H7)
- Samples with excessive or out-of-balance harmonics (generally excessive H3)

The CPM have less experience than the FL10s, and most of them are not flutists. The CPM show divergent preferences from the skilled flutists. In particular, there is a statistically significant difference for samples that the FL10s rated poorly. While FL10s disliked samples with weak upper harmonics, the CPM liked these samples. Further, although not statistically significant, all of the notes that the skilled flutists rated favorably were rated slightly lower by students.

This difference in preference was unexpected. It is unclear if this is indicative of differences in how highly skilled flutists perceive tone quality from the general public.

6.1 Future Work

Acquiring tone samples from additional highly skilled, particularly elite recording artists, could yield valuable insights. Ideally these recording would be acquired in recording studios and administered by skilled sound engineers.

Capturing survey results for a broader set of tone samples would provide more comprehensive information. This would include a wider range of pitches that include vibrato, dynamics, and short phrases.

Further analysis can be conducted with the existing repository of more than 1,600 flute tones. There are opportunities to apply machine learning techniques. These techniques can be applied to not the spectra data, but to the survey results as well.

There are potentially other factors affecting timbre that are outside of the first seven harmonics. For example, descriptors like airy or breathy may be describing background ambient noise, or wind noise from the flutist's air. These types of sounds may be outside of the seven harmonics and not captured by the current HAT implementation.

6.2 Final Words

One of the objectives of this project was to understand if there are quantifiable aspects for "good" tone quality. While there appears to be some correlations, much more work is required towards this goal. The unstated, and underlying question motivating this project was: given some metric for "good" tone quality, can a tool like HAT help musicians improve their tone quality more efficiently? Further research along this avenue could benefit aspiring musicians.