AMA LAB 1: LOW-LEVEL DESCRIPTORS

Antonio Escamilla
Sebastian Vega
UPF- SMC

ABSTRACT

The goal of this lab is to implement a small and simple set of low-level audio descriptors and analyze their distribution over a collection of audio samples.

1. INTRODUCTION

The first part of the lab, aims to study the descriptor values for a set of instrument samples (temporal evolution and global statistics). For this purpose samples from trumpet, cello, harp and flute single notes recordings were selected. Then the values of global features, for all the instrumental samples are analyzed, having in mind how they represent percussive/non-percussive or harmonic/non-harmonic instruments as a way of classifying them. For this particular case 5 instruments were selected and 4 samples within each type of instrument were used, except for the percussive sound, where only one sample was used. Finally 2-D plots (feature space) visualizing the values of 2 descriptors for the different samples are shown and analyzed.

2. INSTRUMENT DESCRIPTORS VALUES

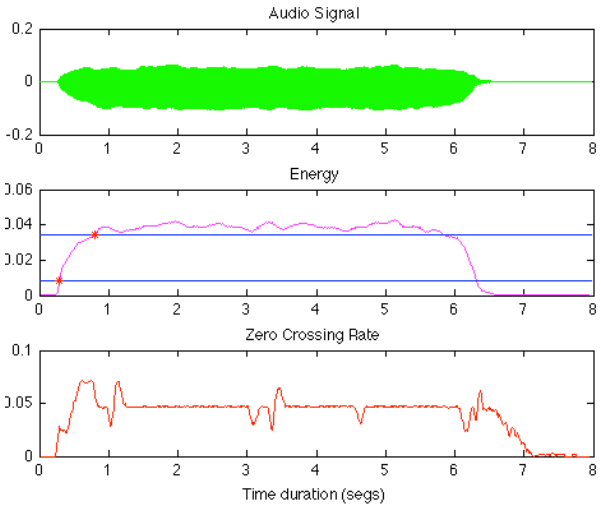

Figure 1. Descriptors in time domain for the trumpet sample 'trpu_gref_mf_do4_12.wav'.

The first instrument subject of analysis is a trumpet sample executed without any articulation or particular technique, being a do4. Depicted in figure1 are: the time signal wave, the energy envelope and the zero crossing rate, in that order. As a way of visualizing the parameters used for the calculation of the log attack time, the 2 energy points used are marked on the second graph.

From the energy envelope one can see that the nature of the sound it's not percussive but also that the attack is long, followed by an even longer sustain. Looking at the zero crossing rate as an indication of the oscillations of the wave on each time frame, one can say that the performance of the note wasn't totally frequency stable, especially at the beginning and end of the sound, probably a product of the player's execution.

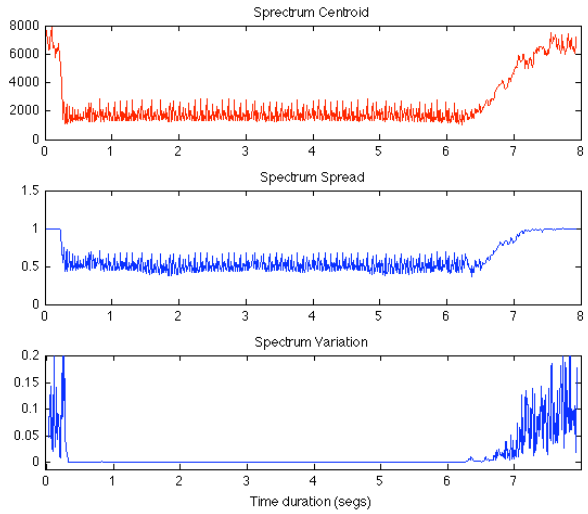

Figure 2. Time evolution of freq. domain descriptors for the trumpet sample 'trpu_gref_mf_do4_12.wav'.

Figure 2 shows three descriptors based on the spectral content of the signal. First, the Spectral Centroid that can be seen as a measure of the center of gravity of the spectrum, taking the frequencies as a distribution with probabilities given by the amplitudes. In the upper plot one can see that the Centroid is changing more or less periodically around a frequency of 2 kHz, but without

making relevant big excursions. This is supported by the Spectrum Spread, the second plot in figure 2, that indicates how the energy is distributed in the frequencies around the Spectrum Centroid, and which presents the same time evolution (small oscillation) as the Spectral Centroid. Finally, the third graph indicates the temporal evolution of how much the spectrum is varying, having two extreme fluctuations on the boundaries of the audio file, caused by the low SNR. That contrasts with the flatness and low values of the middle zone of the plot, where the sustain of the note is apparent.

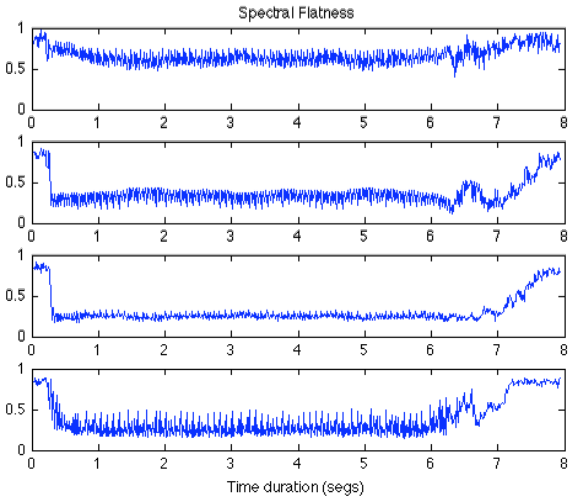

Figure 3. Spectral flatness for the trumpet sample. Freq. bands used(250/500/1K/2K/4K)Hz respectively.

The Spectral Flatness was calculated in four frequency bands as depicted in figure 3. The descriptor is intended to offer a measure of how tonal a sound is by indicating how present are all the frequencies in the range of analysis. For the middle bands (500/1K/2K) Hz, the descriptor indicates high tonal content, as result of low values for the flatness of the spectrum. In the top plot, the Flatness is higher, what means that the audio file has a lower band (250/500) Hz full of frequencies that contribute with energy to the overall perception of the sound. Finally the higher band (2K/4K) Hz, shows again a low floor value of the flatness but with higher oscillation.

2.1 Other Instruments Samples

The following figures, 4 to 6, show the same descriptors mentioned in the previous section, but for a different sound: a pizzicato do3 played on a cello.

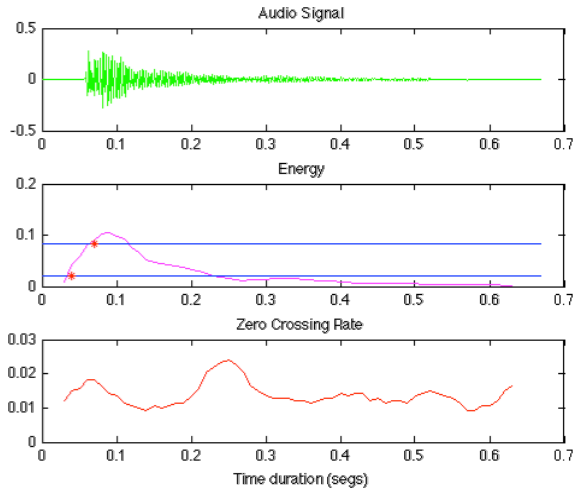

Figure 4. Descriptors in time domain for the cello sample 'vcl_a_pizzsec_mf_do3_12.wav'.

As it is expected, the technique used adds a percussive character to the cello note as it is depicted in figure 4. The energy envelope shows that the sound has a fast attack, a short sustain and a large release with some reverberation tail extending it. Because of this long low level release that is comparable with the noise level, the zero crossing rate is meaningless the second half of the time evolution.

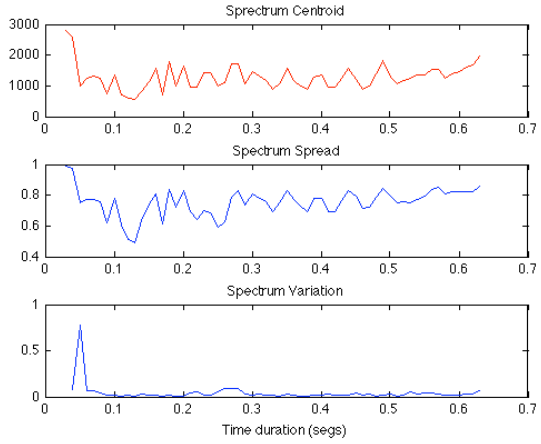

Figure 5. Time evolution of freq. domain descriptors for the cello sample 'vcl_a_pizzsec_mf_do3_12.wav'.

The Spectral descriptors extracted for the cello are shown in fugure 5. The Spectrum Centroid seems to be somehow between 500 and 2K Hz but with a Spread that is varying a lot, making it difficult to extract some information from them. In contrast, the Spectral Variation clearly shows a peak in the attack transient, but then it reach an “average” stable value, but not as flat as the Spectral Variation of the previous sample. Figure 6 shows the Spectral Flatness of the cello.

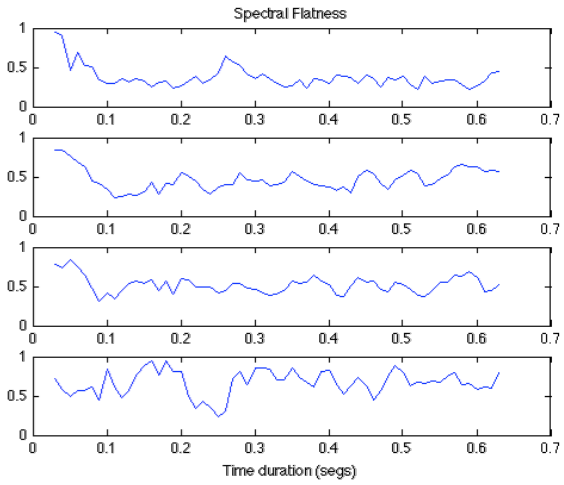

Figure 6. Spectral flatness for the cello sample. Freq. bands used(250/500/1K/2K/4K)Hz respectively.

A plucked harp is now analyzed and from the behavior of the descriptors, which can be seen in figures 7, 8 & 9, a relation between both of them given by the technique used to play, can be observed.

Contrary to bowed or wind instruments where the player can apply energy for a long time to a note performance, therefore extending its sustain; we observe in figure 7, when a string is plucked, how the energy envelope drops fast after its maximum value is reached. This causes: log attack values very much negatives, but also very short fragments in time where the zero crossing rate stays steady; both observations are seen in the harp and cello-pizzicato descriptors and will be, for sure, useful when trying to recognize these type of instruments against bowed or wind instruments.

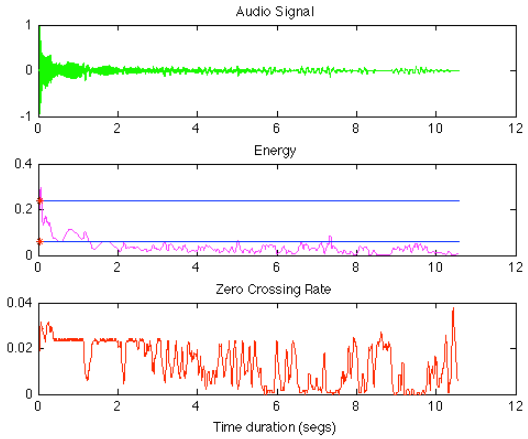

Figure 7. Descriptors in time domain for the harp sample 'harp_gref_mf_do4_12.wav'.

In figure 8 we observe again common characteristics between the harp and the cello-pizzicato samples. This time the fluctuations of the Spectral Centroid evolution over time are much more prominent and in some sense more random. The Spectral Spread around the more undefined Centroid it is always above 0.5 and almost close to 1 from one third to the end of the time. The Spectrum Variation results are consequence of the large decay of the sample, where the low level signal from the instrument mixes together with the background noise, which now becomes comparable in magnitude.

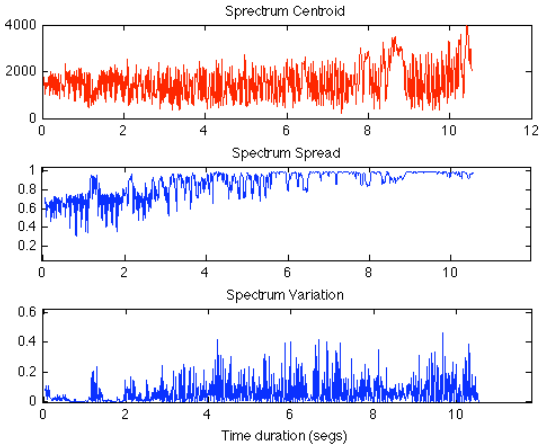

Figure 8. Time evolution of freq. domain descriptors for the harp sample 'harp_gref_mf_do4_12.wav'.

This last seen phenomenon is not really a problem, it's just a characteristic found in large decay instruments that actually comes handy when trying to classify them. In figure 9, the graphs of the Spectral Flatness show low evidence of sinusoidal behavior, especially in the mid-high and high bands, where in fact the plots are constantly reaching a value of 1. Accordingly to the Spectral Flatness results, the tonality of the sound is in the band between 500 and 1K Hz, where a low value is plotted for the first half of the time evolution.

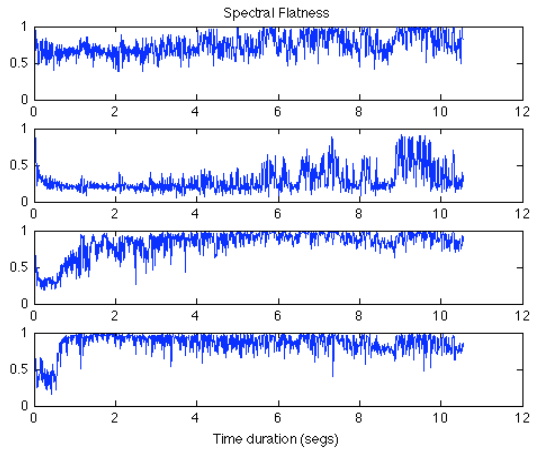

Figure 9. Spectral flatness for the harp sample. Freq. bands used(250/500/1K/2K/4K)Hz respectively.

One last example is depicted in figures 10 to 12, this time a flute sample. Here all descriptors seem to be behaving well or at least in a controlled way. In figure

10 the energy envelope as well as the zero crossing rate shows a steady evolution of each descriptor as long as the sustain last, again a characteristic proper of wind instruments that will help us to identify them.

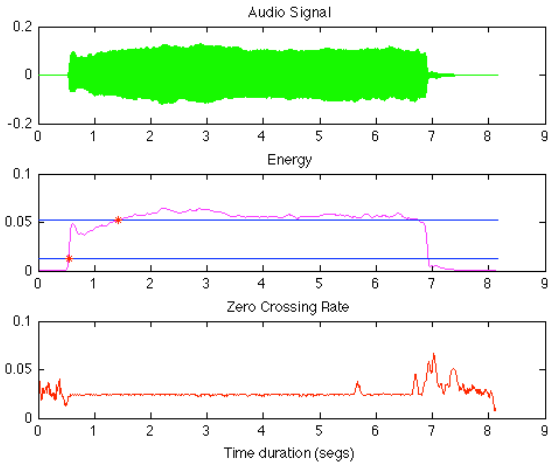

Figure10. Descriptors in time domain for the flute sample 'fltu_gref_mf_do4_12.wav'.

From the figures 11 and 12, we can make a comparison between how similar the spectral descriptors are for the analyzed samples of a trumpet (see figures 2 & 3) and the flute we are talking about.

The Spectrum Centroid is clearly around the same frequency values, which can be understood given that both samples are performing the same exact note, a do4. But apart from that, the Spectral Spread are quite similar showing also the same kind of oscillation during the sustain of the note, and an almost zero stable and steady value for the spectral variation.

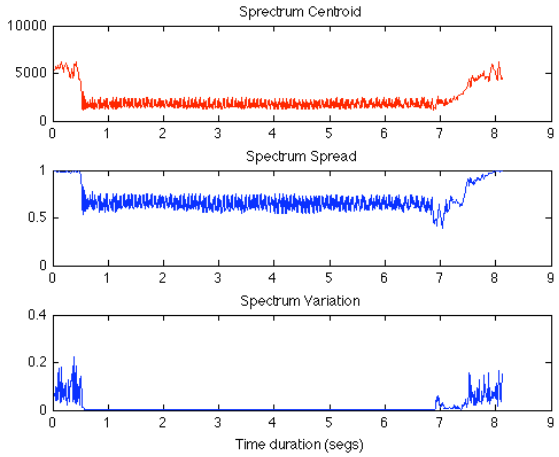

Figure 11. Time evolution of freq. domain descriptors for the flute sample.

If we compare the Spectral Flatness, the similarities are obvious; therefore most of the comments of figure 3, regarding the trumpet sample can be applied for the flute sample. One conclusion of this similarity is that for a classifier that would separate this 2 types of instruments, the Spectral descriptors won't be helpful given the observed behavior between both samples; for the case of a more successful classifier, time domain descriptors as the zero crossing rate, log attack time and temporal centroid will work better.

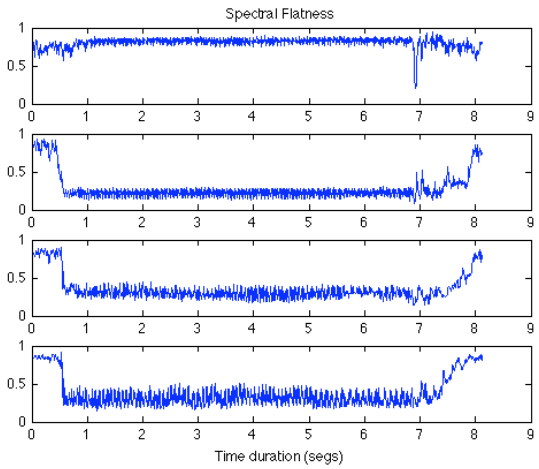

Figure 12. Spectral flatness for the flute sample. Freq. bands used(250/500/1K/2K/4K)Hz respectively.

The table 1, gather the information of the four previous instrument samples and the global descriptors calculated in time and frequency domain.

	Trumpet	Cello-pz	Harp	Flute
Log Attack Time	-0.2924	-1.5229	-2	-0.055
Temporal centroid	3.3229	0.2032	3.6630	3.7924
ZCR mean	0.0388	0.0113	0.0134	0.0260
ZCR std	0.0180	0.0051	0.0089	0.0064
Sp Centr mean	2479.7	2097.7	1536.8	2241.8
Sp Sprea mean	0.5929	0.8589	0.8694	0.6897
Sp Variat mean	0.0146	0.0504	0.0595	0.0126
Sp Centr std	1681.5	913.49	681.18	1202.7
Sp Sprea std	0.1832	0.1222	0.1420	0.1256
Sp Variat std	0.0420	0.0747	0.0782	0.0312

Table 1. Global Descriptors for different instruments.

In the above table we can see that the plucked harp, along with the cello pizzicato have the fastest attack, these instruments have a similar production mechanism and that is why they share this property. In the next sections we will see that other descriptors and other sound samples support this.

Next section studies the distribution of descriptors for different groups of instrument samples, then we also explore the visualization of feature spaces, which are two dimensional plots that allow us to find patterns easily.

3. DESCRIPTOR VALUES FOR GROUPS OF INSTRUMENTS

The aim of this section is to analyze the distribution of descriptor values for different instrument groupings. For this purpose several samples were selected and grouped together depending on their sonic nature. Five different groups were created, each with four samples, except group number five which only contains one sample. The groups are organized as follows. Samples 1-4 are all flute sounds, samples 5-8 are guitars, samples 9-12 are trombones, samples 13-16 are cellos, and finally sample 18 is a cello note played pizzicato style (therefore it is labeled as percussive). In order to allow the comparison between descriptor values of several samples we need to calculate global values for each sound. Except from the log attack time and the temporal centroid all descriptors in this work are calculated in a frame wise manner. Therefore in order to obtain the global values we take the mean and variance of each descriptor's time series. For the two exceptions mentioned above there is no need to do this as they are scalar values. The figure below shows the distribution of values for the temporal centroid and log attack time.

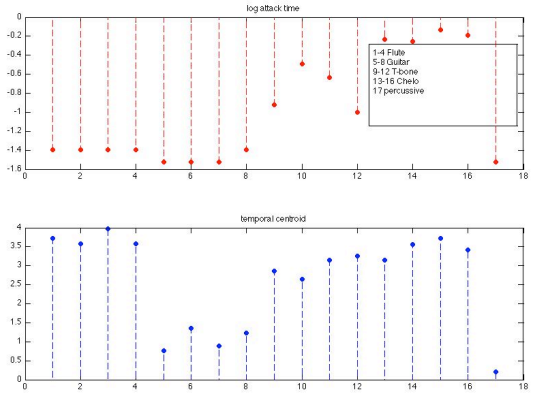

Figure 1 - Distribution of values for log attack time and temporal centroid descriptors

It is possible to arrive at some conclusions by analyzing the distribution of results. Namely, the guitars and pizzicato cello have the fastest attacks; the flutes show a slightly slower attack than these, followed by the cellos and trombones, which show the slowest attacks. All of these results can be used to gain some insight into the mechanisms used to produce each sound. For example, the guitars and pizzicato cello have the fastest attacks, in both of these the fingers are used to pluck the strings. In the case of the cello, it is intuitive that a bow will take some time to displace along the string so the attack of these types of sound are somewhat slower. The results of the temporal centroid also agree with intuition, the longer the sound (more sustain) the larger the value of the temporal centroid. This is why plucked guitars and pizzicato cello have the lowest values, while longer sounds like sustained flute notes, cello notes and sustained trombones have larger values.

3.1. Frame Based Descriptors

As already mentioned, for this descriptors we need to show some simple statistical quantities as their global values. The mean and variance of each descriptor's time series is calculated for every sample and this allows a simplified comparison between different groups. The figures below illustrate this.

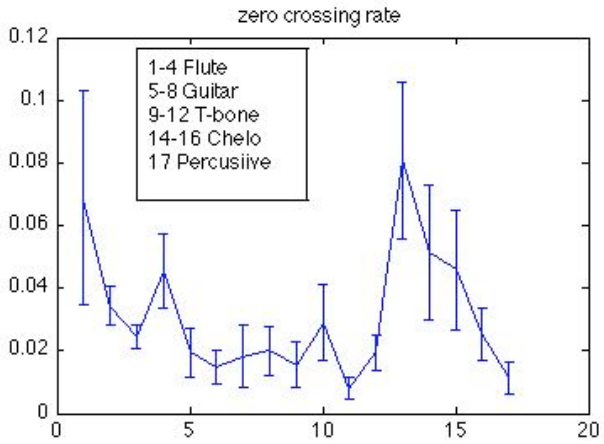

Figure 2 - Zero crossing rate mean and variance

The zero crossing rate shows a similar result for all guitar samples, the mean, as well as the variance is pretty uniform for this group. The trombones also seem to show low uniform values, this hints at the perfectly periodic nature of trombones. In the other hand, the flute seems to show random distribution of values as well as the cello. It is difficult to come to conclusions with this descriptor because we are not considering any noisy samples; also, the zero crossing rates might be affected by pitch and depends on the recording conditions of each sample (SNR). If our sample set included a group of cymbals for example, or any other noisy, non periodic sound, then the zero crossing rate would show an area of high values for this group. Also the values of the zero crossing rates could be energy weighted so that areas of low amplitude in the

waveform are not overtaken by the noise of the recording device and thus it would yield more accurate results.

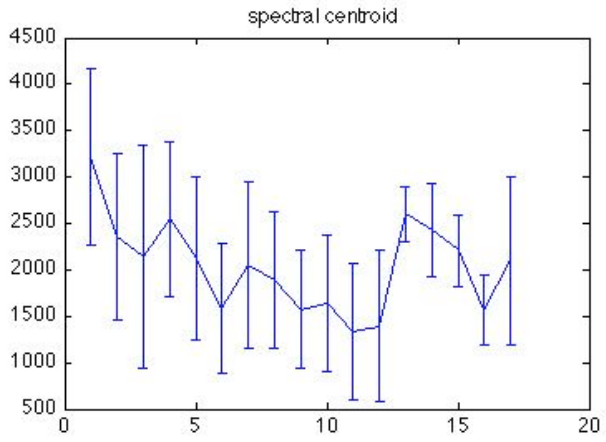

Figure 3 - Spectral Centroid mean and variance

The spectral centroid is also affected by pitch and playing style, but we can observe that in general flutes seem to have a higher spectral centroid value, this is perceptually relevant as flutes do produce bright sounds most of the time. The variances are very similar for most instruments except the cello, which shows lower variances. We can also see that trombones show a pretty uniform distribution both in mean and variance, which could mean that these instruments have a consistent timbre.

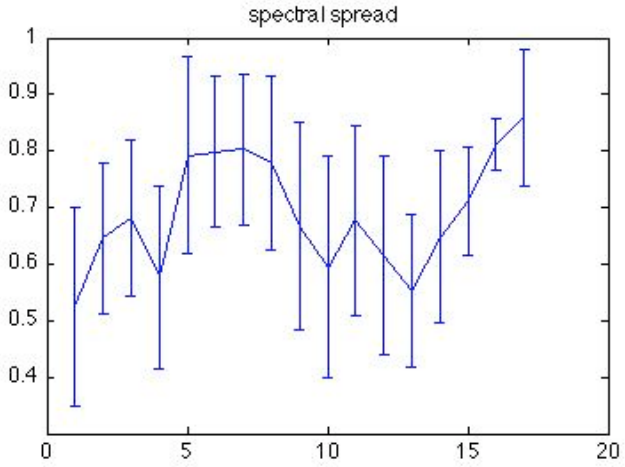

Figure 4 - Spectral Spread mean and variance

There are two things that stand out in the spectral spread. The guitars and the pizzicato cello seem to have the highest value of it. This could be linked to the fact that they are the most transient sounds in the set (due to the finger plucking). The high value of spectral spread might be due to the fact that the decays of these sounds are somewhat longer than the decays of other instruments, and can be affected by room acoustics.

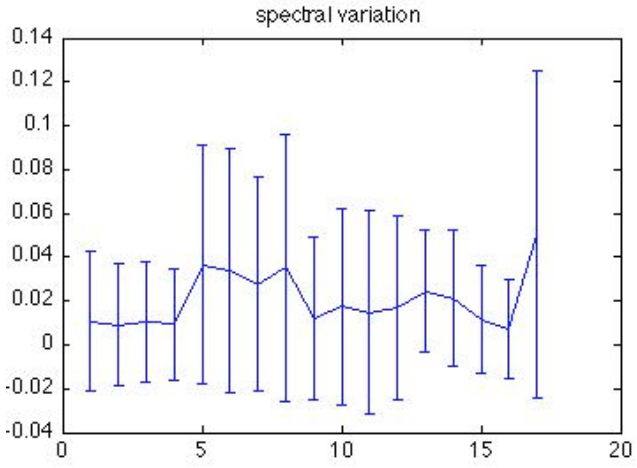

Figure 5 - Spectral variation mean and variance

The spectral variation shows very uniform results across groups. Again the guitars and pizzicato cello have the highest value and it could be due to the reasons discussed above. The flutes and trombones show very uniform results across samples, which again, could mean that these instruments have a consistent timbre. The last of the frame based descriptors is the spectral flatness which is calculated for different frequency bands as already mentioned.

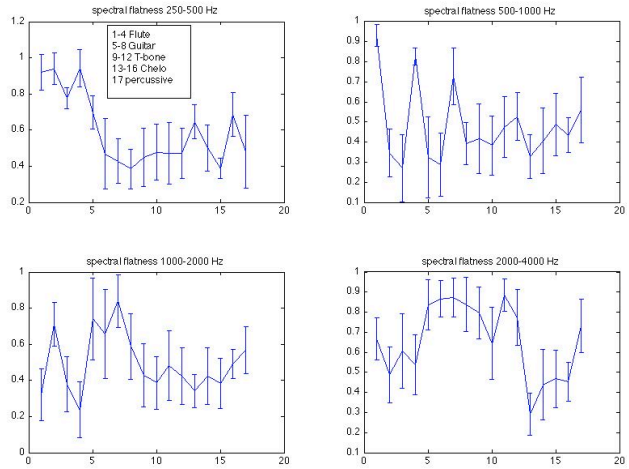

Figure 6 - Spectral flatness mean and variance

As in the case of zero crossing rate it is difficult to interpret these results without making assumptions. None of the samples in the instrument set was particularly noisy in nature and the results seem to be varying a lot within each group. In general, the trombone shows low values of the flatness measure, which indicates high tonality, except for the higher frequency band, which shows a clear increase. This increase in the higher band might be due to

the noise that is produced by the musician's blowing into the instrument. That concludes this section, it is evident that there are some descriptors that are able to characterize certain sonic properties, and it is also evident that sometimes descriptors are not able to characterize much. When using descriptors to analyze sounds it is better to have a particular characteristic in mind and to understand which descriptor could be useful rather than testing all descriptors randomly to try and find patterns.

4. TWO-DIMENSIONAL VISUALIZATION

The use of 2-D plots to visualize a feature space is extremely useful as it aids in the localization of patterns. When looking at the feature space sometimes it is evident that different groups cluster together. This information can give us cues for the classification of different classes of sounds. A clear example of this is shown below, where the log attack time and temporal centroid are useful in separating these samples by group.

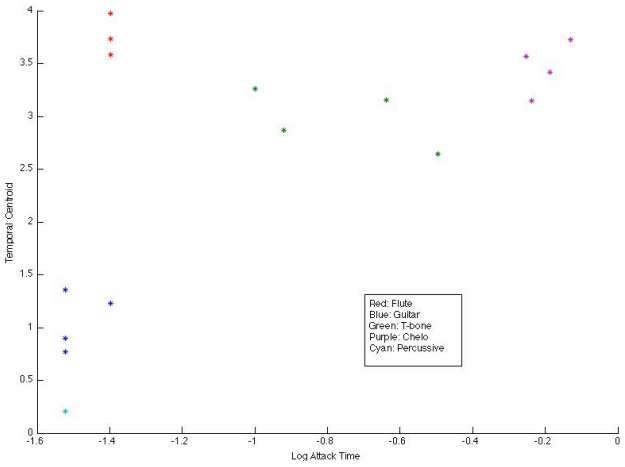

Figure 7 - Temporal centroid and log attack time feature space. The different groups are clearly clustered

In the example above all of the different groups are clustered together, this type of distribution would allow the proper classification of the different groups with a high accuracy. Figure 8 shows another feature space for the spectral variation and zero crossing rate descriptors. This feature space also shows a tendency to cluster the different groups, although there are some points that do not precisely do this. If we used this feature space for classification purposes we could get some errors, but nothing extreme, in general it is still a good distribution.

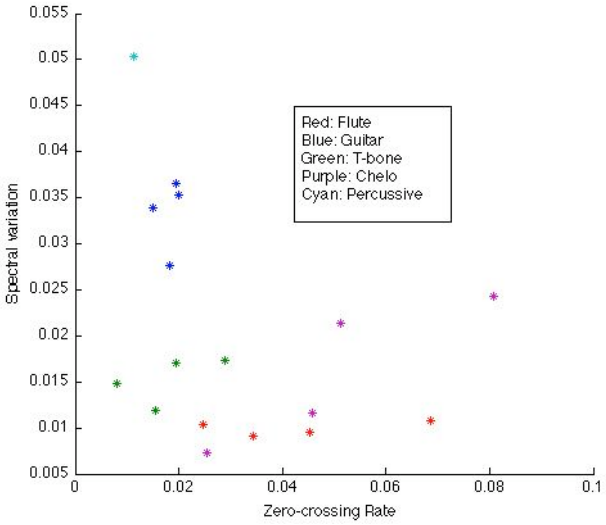

Figure 8 - Zero crossing rate and spectral variation feature space. Still shows some clustering of the groups

A similar case is the feature space depicted in figure 9. Most of the groups are clearly clustered but some points (especially the cellos) seem to be misplaced. Maybe with a complex classification function it could still be possible to obtain good classification results.

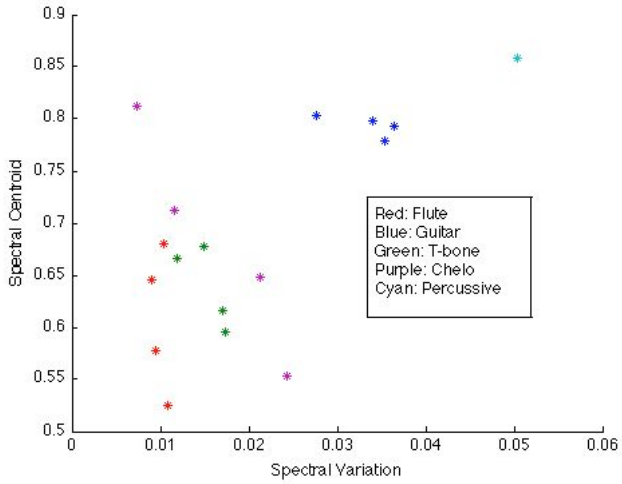

Figure 9 - Spectral variation and spectral centroid feature space. Still some clustering is observable.

Finally, the last feature space shows the same problem. Some cello samples seem to be misplaced and the rest of the groups are properly clustered (especially the guitars).

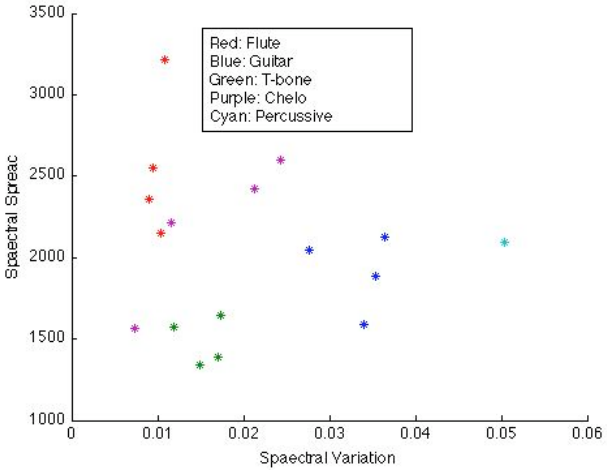

Figure 10 - Spectral spread and spectral variation feature space.

After observing the feature spaces for different combinations of descriptors it can be concluded that the temporal centroid/log attack time combination is the best for separating between these types of sounds. Also, the cello seems to show the most problems in the rest of the cases as it has some random samples in an otherwise perfectly clustered space. In general the results were good and some patterns can be clearly recognized.

5. REFERENCES

-Perfecto Herrera-Boyer, Anssi Klapuri and Manuel Davy, *Automatic classification of pitched musical instrument sounds*, 2006.

-Geoffroy Peeters, *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*, 2004.

-MPEG-7 Multimedia Software Resources, http://mpeg7.doc.gold.ac.uk/