PERCEPTUALLY BASED PITCH SCALES IN CEPSTRAL TECHNIQUES FOR PERCUSSIVE TIMBRE IDENTIFICATION

William Brent

Department of Music and Center for Research in Computing and the Arts, UCSD

ABSTRACT

Different types of cepstral analysis are compared in the context of a percussion instrument classification external for Pd. For raw cepstrum, mel frequency cepstrum, DCT-based cepstrum, and bark frequency cepstrum, various parameter settings are applied to a standardized test. Significant score improvement can be seen when moving from cepstrum to mel cepstrum, and further improvement is achieved using bark cepstrum. Considering the prevalence of MFCCs as a feature vector for timbre, it is suggested that BFCCs are at least as effective, and more appropriate in light of accompanying psychoacoustic research.

1. INTRODUCTION

Classically, spectral techniques making use of a short-time Fourier transform have been the dominant solution for timbre classification. A general problem with spectral domain methods is high dimensionality. The solution to this problem is to forgo some spectral resolution in order to reduce total points of comparison. For example, a great number of techniques are currently in use as MPEG7 audio descriptors [8]; some popular examples that reduce data size are spectral flux, spectral flatness, spectral rolloff, spectral centroid, spectral smoothing, and cepstral analysis.

In [4], a timbre classification system is described that smooths spectra using a bank of eleven filters composed so that there are two filters per octave. Assuming that the instruments in question have unique distributions of energy in relation to the bands of the filterbank, it is possible to accurately distinguish between timbres by creating training-based templates for comparison against incoming signals. But information other than general spectral envelope—such as a strong pitch component in a specific instance of an instrument articulation—will also be reflected under this technique. In some situations, it would be ideal to have an analysis method that identifies, for instance, a timpani without regard to its tuning.

The final spectral domain technique mentioned above, cepstral analysis, is often presented as such a method. This paper will evaluate the technique by way of documenting the development of a Pd external for classifying percussion instruments. In addition to straight cepstral analysis, two perceptually biased versions—the mel-frequency and bark-frequency cepstrum—will be introduced. The Pd implementations of all three techniques will be compared in order to explore the effects of perceptual frequency scales in timbre classification.

2. CEPSTRUM

The real cepstrum (x_{RC}) is defined as

$$x_{RC}(n) = \Re[\, IFT\{ln|X(k)|\}\,] \tag{1}$$

where $X(k)$ is the frequency domain representation of a signal $x(n)$, and $\Re$ denotes the real portion of the inverse Fourier transform. Functionally, cepstrum∼ is very similar to the classification mode of bonk∼. Both require that the user give training examples of the percussion instruments that are to be identified, whose analyses are stored as templates. Once training is complete, any new incoming signals are compared against the stored templates, and the nearest match is output as the index number of the appropriate instrument as assigned during training. The nearest match is identified based on Euclidean distances between the cepstrum of an incoming signal and stored template cepstra. Unlike bonk∼, cepstrum∼ does not have an attack detection mode; it simply takes a cepstral snapshot when it receives a training or identification request. This means that the length of time between a detected onset (as reported by bonk∼) and cepstral analysis can be chosen.

2.1. A Standardized Test

The collection of instruments that fall under the category of percussion is vast. The process of choosing a suitable set of instruments for this test was guided by three desires: diversity of material, diversity of spectrum, and relatively short decay. The chosen training set includes a low tom, wooden plank, Chinese cymbal, nipple gong, cabaça (shaker), metal bowl, bongo, small anvil, tambourine, thundersheet, conga, and wooden box. The test sequence consisted of one strike of each instrument in the order given above, at a tempo of roughly 108 bpm. 10 runs were recorded, followed by 3 additional runs at roughly 180 bpm. With a total of 13 runs through the 12 instruments, the complete test consists of classifying 156 attacks.

2.2. cepstrum~

Figure 1 shows a series of plots from the initial testing process using a 1024-point analysis window and based on 5 training examples for each instrument. It begins with post-onset analysis time (AT) set to 0 ms, with each subsequent plot along the y axis (moving away from the reader and to the left) representing analyses taken for AT values that increase in 1 ms increments. The entire image shows AT settings from 0-35 ms. Along the x axis (moving away from the reader and to the right), the effects of incrementing the cepstral coefficient range (CCR) setting can be seen, showing scores resulting from using a range of 0-50, 0-250, and 0-500 cepstral coefficients. The z axis (vertical) shows normalized scores that fall between 0 and 1, where 1 represents a perfect score of 156 accurate classifications. Note that the actual range of the z axis is about 0.7-1, as the lowest scores were near 70%.

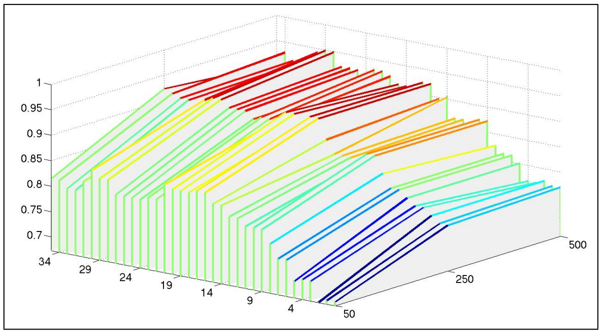

Figure 1. cepstrum~ scores.

The plot reveals a few interesting trends. Increasing CCR to 0-250 clearly improves results at all AT settings. Subsequent increases in the CCR to 0-250 do not dramatically improve or decrease scores. There is a wide area of stable 90%+ accuracy (red) between AT=16-35 ms for higher CCR values. This plateau begins early along the AT axis, which is desirable in terms of reducing latency. A clear tradeoff for the AT parameter can be seen: waiting longer to take a snapshot for analysis presumably captures more relevant acoustic information and generates higher scores. Regarding CCR, there seems to be no reason to use a range smaller than 0-250, where scores are clearly lower in general.

While promising, cepstrum~'s performance is limited to ~90% accuracy for total latency times under 30 ms in a relatively simple test. Other methods could yield higher accuracy with less latency. Cepstral analysis is based entirely on objective measurements of sound, and does not make use of scales related to human pitch perception that have been constructed based on the systematic tracking of subjective judgements. Drawing on a perceptually-based frequency scale, a popular form of cepstral analysis for feature detection is the mel-frequency cepstrum [2] [1].

3. MEL CEPSTRUM

In 1937, a perceptual scale for measuring pitch was proposed in [6]. Based on the experimental data of 5 subjects, the authors hoped to discover a frequency unit that could be manipulated arithmetically yet remain observationally verifiable. In reference to melody, this unit was named the mel. For any particular mel value, one should be able to double it, then convert both the original and doubled values back to a frequency scale and confirm through experiment that the doubled mel frequency is judged to be twice as high in terms of pitch. Likewise, halving or tripling a mel value should lead to appropriately scaled perceptual results. The general formula for calculating mels is

$$Mel(f) = 2595 \times log_{10}(1 + \frac{f}{700}) \tag{2}$$

where f is frequency in Hz.

3.1. Mel Frequency Cepstral Coefficients

The process for computing Mel Frequency Cepstral Coefficients (MFCCs) differs from raw cepstrum computation considerably. It requires a bank of overlapping triangular bandpass filters evenly spaced on the mel scale, and the final transform is a discrete cosine transform (DCT) rather than a DFT. MFCCs are defined mathematically as

$$MFCC_i = \sum_{k=1}^{N} X_k \, cos[i(k - \frac{1}{2})\frac{\pi}{N}]; \;\; i = 1, 2, \ldots, M \tag{3}$$

where M is the number of desired cepstral coefficients, N is the number of filters, and X_k is the log power output of the kth filter. Mel scaling and smoothing significantly reduces the size of spectral envelope data. The extent of reduction depends on sampling rate, window size, and the mel spacing of the filterbank. Using equation (2), the Nyquist Frequency at a sampling rate of 44100 is calculated as 3923 mels. An even spacing of 150 mels would produce 27 mel values below Nyquist, which correspond to 25 overlapping filters. Multiplying the log power spectrum against this filterbank compresses the first 512 bins of a 1024 point analysis window into a smoothed 25 point estimation with a weighting based on the mel scale.

The DCT at the end of the MFCC computation is the other fundamental difference from raw cepstrum. It is a significant enough change to warrant exclusive testing, which will be covered at the end of this section. [2] proposes that the DCT approximates decorrelation obtained through Principal Component Analysis (PCA). If the mel-frequency cepstrum yields higher scores than raw cepstrum, credit cannot be assigned to the mel scale's effectiveness until the DCT step has been tested with spectra that have been smoothed according to a linear scale. The exclusive effects of the mel scale will then be clear.

3.2. mfcc∼

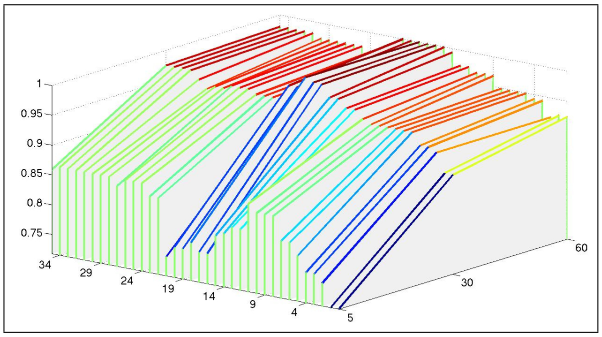

Figure 2. mfcc∼ scores.

The filterbank used in mfcc∼ introduces at least one new parameter for testing: the mel spacing between filters. As mel spacing (MS) becomes narrower, more filters will fit beneath the Nyquist Frequency, directly affecting the CCR parameter. The test below was carried out with MS=60, which expands the potential CCR to 0-63. An additional parameter worth considering is the method for representing the power spectrum under each filter. This implementation uses the normalized fourth root of total power. Omitting normalization significantly reduced scores, but the choice of total or mean power had no impact. Figure 2 shows test scores for MS=60 up to AT=35, and CCR=0-60.

Compared to cepstrum∼ results in Figure 1, accuracy at AT=0 ms has improved from ∼75% to ∼90%+. A jump above 95% accuracy occurs only a few milliseconds later at AT=5. Already, it can be seen that the MFCC technique provides higher accuracy at lower latency. A plateau of 100% accuracy is found between AT=15-20 ms at CCR values of 0-60, only to dip down to ∼95%+ for further AT settings. Following values along the CCR axis reveals that increasing CCR beyond CCR=0-30 does not drastically affect accuracy. However, since the mel scaling and smoothing step reduced the number of analysis points considerably, there is no apparent reason to use anything but the complete set of MFCCs when calculating Euclidean distance for template comparison.

Having seen the effectiveness of the MFCC technique as a whole, we can evaluate the role of the mel scale in particular. Based on the number of filters in a filterbank with 60 mel spacing, we will use a linearly spaced filterbank for the spectral smoothing step before computing the DCT. If results are closer to the scores from cepstrum∼, there will be more reason to believe that the characteristics of the mel scale are responsible for the significant score improvement seen above.

Although the width of the first few filters in the 60 mel spaced filterbank is only about 40 Hz, a linear filterbank at that spacing is impractical for a 1024 window with 43 Hz bin spacing. A larger spacing of 300 Hz produces 72 filters below Nyquist, which is slightly more than the number of filters in a 60 mel spaced filterbank. Filter widths above 300 Hz do not occur in the mel spaced filterbank until around 5 kHz, so the loss of resolution in the low end of the spectrum is indeed the point of comparison between these two scales. Figure 3 shows score results.

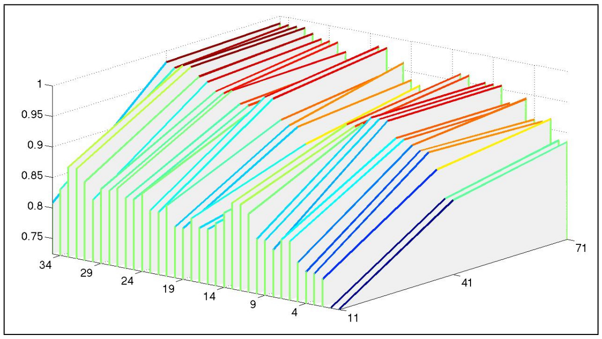

Figure 3. DCT-based cepstrum scores.

Scores without mel weighting are lower in general and never reach 100%, but are significantly higher than those of the raw cepstrum, indicating that the DCT may be partly responsible for the improved scores shown in Figure 2. For this test, the specific contribution of the mel scale (i.e., its property of favoring lower frequencies) is now clear. At this point, we can evaluate whether or not a frequency curve based on the more extensively researched critical bandwidths is equally effective.

4. CRITICAL BANDS AND THE BARK SCALE

Critical bands refer to frequency ranges corresponding to regions of the basilar membrane that are excited when stimulated by specific frequencies. An overview of multiple experiments establishing the boundary and center frequencies of critical bands is given in [9]. Critical band boundaries are not fixed according to frequency, but dependent upon specific stimuli. Relative bandwidths are more stable, and repeated experiments have found consistent results. In frequency, these widths remain more or less constant at 100 Hz for center frequencies up to ∼500 Hz, and are proportional to higher center frequencies by a factor of 0.2. In 1960, Zwicker introduced the Bark as a unit based on critical band boundaries, named after the inventor of the unit of loudness level: Barkhausen.

Unlike the mel scale, the Bark unit stands upon a large foundation of evidence. As Zwicker et al. put it, the "critical band has the advantage . . . that it does not rest on assumptions or definitions, but is empirically determined by at least four kinds of independent experiments." [9] The four unique strategies for locating critical band boundaries that they refer to are threshold, masking, phase, and loudness summation.

Bark-frequency cepstral analysis has been applied elsewhere [3], but is quite rare in comparison to mel-frequency cepstrum. Despite a difference in terms of verification by independent experiments, several sources note that Barks relate very strongly to mels [9] [5], the rough guide being that multiplying Barks by 100 produces a curve similar to the mel scale. Only a handful of mel values were determined directly by experiment; intermediate values are projected based on equation (2). Likewise for Barks, since there are a fixed number of critical bands that correspond to the 24 Barks, values at arbitrary subdivisions between boundaries or beyond the 24th Bark must also be calculated with a general formula. Equation (4), taken from [7], will be used here, where f is frequency in Hz.

$$Bark = [26.81 \times \frac{f}{(1960+f)}] - 0.53 \tag{4}$$

4.1. Bark Frequency Cepstrum and bfcc∼

Implementation of Bark weighting in place of mels is straightforward. The collection of frequencies used for filterbank construction will merely be generated based on equation (4) rather than (2). A Bark spacing parameter (BS) functions identically to the MS parameter of mfcc∼. Figure 4 shows results for AT=0 through AT=35, and CCR=0-6, CCR=0-26, and CCR=0-46 (the total available coefficients produced from half-Bark spacing).

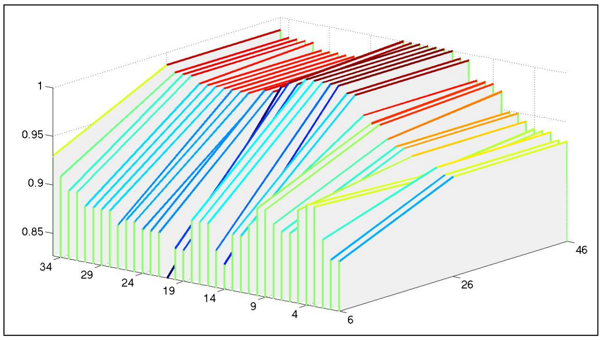

Figure 4. bfcc∼ scores.

The lowest score (83%) at AT=21 ms is higher than the lowest score from the mfcc∼ test (72%) when using MS=60. As in the mfcc∼ test, a plateau of 100% accuracy exists, this time beginning earlier at AT=14 ms instead of 15 ms, and extending to 20 ms. The improvement is slight, but the fact that the plateau starts 1 ms earlier and is 1 ms wider is certainly a strength. Like mfcc∼ performance discussed above, AT values before 14 ms produce consistent and useful results above 92%.

At the highest CCR settings, bfcc∼ scores are higher for the first three AT settings, lower for AT=3-5ms, and consistently higher from AT=8 ms onwards. We can conclude that Bark units are at least as useful as mels for weighting a spectrum, and possibly more appropriate. As the Bark scale can produce slightly more accurate results and has a larger body of research behind it, perhaps the widespread use of mel weighting in cepstral analysis should be questioned.

5. DISCUSSION

In this report, we have directly seen improvements that can be gained by using two types of perceptual scales; however, we cannot necessarily conclude that the increased accuracy should be attributed to the value of perceptual information. From an objective and appropriately skeptical standpoint, we have merely seen that an emphasis on lower spectral content improves results. Only through further experimentation can we become confident that such improvements are not partially coincidental.

6. REFERENCES

[1] S. Dubnov, G. Assayag, and A. Cont, "Audio oracle: A new algorithm for fast learning of audio structures," in *Proceedings of the International Computer Music Conference*, Copenhagen, Denmark, 2007.

[2] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proceedings of the International Symposium on Music Information Retrieval*, 2000.

[3] F. Mörchen, A. Ultsch, M. Thies, I. Löhken, M. Nöcker, C. Stamm, N. Efthymiou, and M. Kümmerer, "Musicminer: Visualizing timbre distances of music as topographical maps," CS Department, Philipps-University Marburg, Germany, Tech. Rep. 47, 2005.

[4] M. Puckette, T. Apel, and D. Zicarelli, "Real-time audio analysis tools for pd and msp," in *Proceedings of the International Computer Music Conference*, 1998, pp. 109–112.

[5] T. Rossing, F. Moore, and P. Wheeler, *The Science of Sound*. New York: Addison Wesley, 2002.

[6] S. Stevens, J. Volkman, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, pp. 185–190, 1937.

[7] H. Traunmüller, "Analytical expressions for the tonotopic sensory scale," *Journal of the Acoustical Society of America*, vol. 88, no. 1, pp. 97–100, 1990.

[8] X. Zhang and Z. Ras, "Analysis of sound features for music timbre recognition," in *Proceedings of the IEEE CS International Conference on Multimedia and Ubiquitous Engineering*, 2007, pp. 3–8.

[9] E. Zwicker, G. Flottorp, and S. Stevens, "Critical bandwidth in loudness summation," *Journal of the Acoustical Society of America*, vol. 29, pp. 548–557, 1957.