PERCEPTUALLY BASED PITCH SCALES IN CEPSTRAL TECHNIQUES
FOR PERCUSSIVE TIMBRE IDENTIFICATION

William Brent
University of California, San Diego
Department of Music and Center for Research and Computing in the Arts
wbrent@ucsd.edu

ABSTRACT

Different types of cepstral analysis are compared in the context of percussion instrument classification externals for pure data (Pd) by the author. For raw cepstrum, mel frequency cepstrum, discrete cosine transform-based cepstrum, and Bark frequency cepstrum, a standardized test is run with various parameter settings. Significant score improvement can be seen when using mel cepstrum rather than raw cepstrum, and further improvement is achieved using Bark cepstrum. Considering the prevalence of mel frequency cepstral coefficients as a feature vector for timbre, it is suggested that Bark frequency cepstral coefficients 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 [7]; some popular examples that reduce data size are spectral flux, spectral flatness, spectral roll-off, spectral centroid, spectral smoothing, and cepstral analysis.

In [3], a timbre classification system is described that performs spectral smoothing 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 pure data (Pd) externals developed by the author for classifying percussion instruments. In addition to an external that employs straight cepstral analysis (cepstrum∼), two perceptually biased versions based on mel-frequency and Bark-frequency cepstrum (mfcc∼ and bfcc∼) will be introduced. The performance of these externals will be compared in order to explore the effects of perceptual frequency scales in percussive timbre classification. Results will be reported in relation to various parameter settings for each external, indicating optimal ranges for practical use in real time classification applications.

2. METHOD

After initial work in MATLAB, these experiments were conducted using Pd with the intention of developing tools for real-time applications. Because immediate analysis results are a priority in order to produce synchronous synthetic sound based on classification, the tests below are based on an analysis of only the attack portion of the sounds studied.

Analysis was performed with a 1024-point window (23 ms at a sampling rate of 44.1 kHz). In order to obtain precise time resolution, analysis windows were overlapped every 64 samples.

The process of choosing a suitable set of instruments 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. A single Shure SM57 cardioid microphone was used as the input source. 20 training examples of each instrument were recorded, but only 5 were used in actual training. The test sequence consisted of one strike of each instrument in the order given above. 10 runs were recorded at a tempo of roughly 108 bpm, 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 unique attacks.
3. CEPSTRUM

The real cepstrum \((x_{RC})\) is defined in [4] as
\[
x_{RC}(n) = \Re \{ \text{IFT} \{ \ln |X(k)| \} \}
\]
where \(X(k)\) is the frequency domain representation of a signal \(x(n)\), and \(\Re\) denotes the real portion of the inverse Fourier transform. The real cepstrum is sometimes equivalently defined as the real part of the forward Fourier transform of the logarithm of magnitude spectrum.

3.1. cepstrum~

Cepstral coefficients are put to use as a feature vector in the cepstrum~ external, which is functionally similar to the classification mode of bonk~.\(^1\) Both require that the user give training examples of the percussion instruments that are to be identified so that their analyses can be stored as a database of 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. 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 exact time of cepstral analysis can be chosen in relation to a detected onset (as reported by bonk~).

Figure 1 shows a series of plots from the testing process using cepstrum~. The plots begin 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. The \(AT\) setting refers to the time at which the analysis window ends relative to an onset report from bonk~.

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 0.7-1.0, as the lowest scores were near 70%.

The plot reveals a few interesting trends. Increasing CCR to 0-250 clearly improves results at all \(AT\) settings. There is a wide area of stable 90%+ accuracy (red) between \(AT=16-35\) ms for higher CCR values. While these results are useful, 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 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 mel-frequency cepstral analysis [1].

4. MEL CEPSTRUM

In 1937, a perceptual scale for measuring pitch was proposed in [5]. 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 linear 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. A general formula for calculating mels is

\[
\text{Mel}(f) = 2595 \times \log_{10}(1 + \frac{f}{700})
\]

where \(f\) is frequency in Hz.

4.1. 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 Fourier transform. MFCCs are defined mathematically as

\[
\text{MFCC}_i = \sum_{k=1}^{N} X_k \cos[(i - \frac{1}{2}) \frac{\pi}{N}]; \quad i = 1, 2, \ldots, M\]

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 \(k^{th}\) filter. Mel scaling and smoothing significantly reduces the size of spectral envelope data and emphasizes lower frequency content.
The use of a DCT is the other fundamental difference from raw cepstrum. It is significant enough to warrant exclusive testing, which will be covered at the end of this section. [1] proposes that the DCT approximates decorrelation obtained through Principal Component Analysis.

4.2. mfcc

The filterbank used in mfcc introduces at least one new parameter for testing: the mel spacing between filters. Though it is common to use a fixed filterbank composed with one specific mel spacing, mfcc constructs filterbanks based on arbitrary user-specified mel spacing. 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. 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 a few milliseconds later at AT=5. Already, it can be seen that for this set of instruments 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.

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 can instead use a linearly spaced filterbank for the spectral smoothing step before computing the DCT. Figure 3 shows score results from this process.

Scores without mel weighting are lower in general and never reach 100%, but are 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 emphasizing lower frequencies) is now apparent.

5. 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] Two of the common strategies for locating critical band boundaries that they refer to are masking and loudness summation.

Participants in an experiment exploring the latter method were asked to match loudness between single tones and multiple tone complexes of varying frequency width \( \Delta f \). Modulating values of \( \Delta f \) within a frequency-dependent critical bandwidth did not affect subjects’ loudness judgements, but increasing \( \Delta f \) beyond this bandwidth resulted in increased loudness. The points at which such loudness increases occurred were correctly predicted according to proposed critical band boundaries.

In the case of masking, studied in [8], a small band of noise is placed between two tones. At very low noise sound pressure levels, the tones mask the noise. As the frequency width between the tones is gradually widened, the sound pressure level at which the noise ceases to be masked remains constant until a particular tone spacing is reached, and the masking ceases at significantly lower levels [9]. When the noise and tones are processed within separate critical
bands, masking effects are decreased. The frequency spacing at which this occurs relates to the critical band.

5.1. Bark Frequency Cepstrum and bfcc

Bark-frequency cepstral analysis has been applied elsewhere [2], 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]. 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 be calculated with a general formula. Equation (4), taken from [6], will be used here, where \( f \) is frequency in Hz.

\[
\text{Bark} = \left[26.81 \times \frac{f}{(1960 + f)}\right] - 0.53 \quad (4)
\]

Implementation of Bark weighting in place of mels is straightforward. The collection of frequencies used for filterbank construction will simply 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 \( \Delta T = 0 \) through \( \Delta T = 35 \), and \( CCR = 0-6 \), \( CCR = 0-26 \), and \( CCR = 0-46 \) (the total available coefficients produced from half-Bark spacing).

![Figure 4. bfcc~ scores.](image)

The lowest score (83%) at \( \Delta T = 21 \) 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 \( \Delta T = 14 \) 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%. 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.

6. CONCLUSIONS

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.

7. REFERENCES


