Novelty Measures as Cues for Temporal Salience in Audio Similarity

Mark Cartwright
Northwestern University
EECS Department
2133 Sheridan Road
Evanston, IL 60201 USA
+1 847 467 5905
mcartwright@u.northwestern.edu

Bryan Pardo
Northwestern University
EECS Department
2133 Sheridan Road
Evanston, IL 60201 USA
+1 847 491 7184
pardo@northwestern.edu

ABSTRACT

Most algorithms for estimating audio similarity either completely disregard time or they treat each moment in time equally. However, many studies over the years have noted several factors that affect how much attention we give to certain sounds or parts of sounds (e.g. loudness, the attack, novelty). These findings suggest that some time segments of audio may be more salient than others when making similarity judgments. We believe that if we could estimate this information, we could improve audio similarity measures. This paper presents the results of a human subject study designed to test the hypothesis that sounds segments with high timbral change are more salient than segments with low timbral change. We then investigate whether we can use this information to improve two audio similarity measures: a "bag-of-frames" approach and a dynamic time warping approach.

Categories and Subject Descriptors

H.5.5 [**Sound and Music Computing**]: Signal analysis, synthesis, and processing; H.3.3 [**Information Search an Retrieval**]: Selection process; H.1.2 [**User/Machine Systems**]: Human Factors

General Terms

Human Factors, Experimentation

Keywords

audio novelty measures, audio similarity, audio temporal salience, query-by-example

1. INTRODUCTION

A query-by-example content-based audio retrieval system is an information retrieval system where the user gives the system an exemplar audio file [8], and the system returns other audio results similar to the exemplar. Such systems can be used for searching sound effects libraries, recommending new music, interactive computer music performance, or even programming complex audio synthesizers.

We are primarily concerned with a system that does content-based audio retrieval on a large corpus of short (a few seconds) synthesized sounds. This would let novice users select synthesizer sounds by providing examples to guide synthesis and selecting sounds from the resulting search. In such an application the ability to properly order the top N choices becomes crucial.

While current machine measures of similarity work reasonably well to a first approximation, correlating fairly well with human ratings of similarity from the macro-perspective (e.g. saxophone vs. ocean waves), they are very noisy, correlating poorly with human ratings of similarity in the micro-perspective (Wurlitzer electric piano from Rhodes electric piano). When there are many similar audio objects, these differences could potentially push the most relevant objects out of the first few pages of results of a search application. In this work, we will study approaches to reduce the amount of noise in machine audio similarity measures so that small differences in these similarity measures are perceptually meaningful and may consequently improve micro-perspective similarity measurements and the ranking of the top N objects.

When computing audio similarity, it is common to first extract a time series of features from the audio. Also, it is common to learn feature weights to minimize the error in machine prediction of human similarity ratings. With a time series of features representing audio, data is segmented not only by feature, but by time. Typical practice is to simply treat all of these time segments equally. However, several studies have shown that that some time segments of audio are more perceptually important than other time segments of audio [25, 12, 27, 15]. We will refer to importance attached to a particular time segment of audio as *temporal salience*. We believe that if we could estimate the salience of audio as a function of the spectro-temporal features, we could incorporate this information into audio similarity measures and increase their correlation with human similarity when comparing relatively similar audio objects.

In Section 2, we present psychoacoustic and neuroscience studies on temporal salience cues in sound. From these studies, we conclude that there are at least three factors that affect the salience a listener attributes to a particular time seg-

MIRUM'12, November 2, 2012, Nara, Japan.

ment: loudness, temporal proximity-to-the-onset (e.g. in the attack), and novelty. We would like to utilize these cues to improve audio similarity measures. However, we first must have a way of estimating these values. There already exist methods for estimating perceptual loudness [21, 10] and the attack segment [13, 17, 22, 3]. However, to our knowledge, there is not an established way to either estimate novelty as a cue for temporal salience or to incorporate it into an audio similarity measurement.

In this paper, we attend to a temporal salience cue (novelty). We aim to make reasonable assumptions to simplify the problem and propose measures for short-term local change in audio. We present a human subject study to evaluate two potential measures for estimating this cue, and we present examples of how to incorporate these measures into methods for audio similarity.

2. BACKGROUND

The idea that some time segments of audio are more important than others is not new in psychoacoustics. We refer to importance attached to a particular time segment of audio as temporal salience. The studies in this section suggest that there are three potential cues for temporal salience in audio: loudness, temporal proximity-to-the-onset (i.e. related to the attack), and novelty.

As common sense establishes and Huron notes [14], loud sounds induce higher levels of arousal due to the fact that loud sounds are more likely to signal danger. Several studies over the past 40 years have reported that humans place importance on the *attack* when performing instrument identification and instrument similarity tasks [25, 12, 27, 15]. Escera [6] noted that humans react and involuntarily attend to sounds that are novel more than sounds that are not, where novelty here refers to when the sound stimulus is "new or relatively rare in relation to the recent history of stimulation" [26]. While Escera focused on a time scale in which they were concerned with the novelty of whole sound objects, we are concerned with comparing individual sound objects to each other. Therefore, we are concerned with the novelty of events *within* a sound object. We are unaware of any studies that have investigated this phenomenon. We therefore postulate that the perceptual phenomenon reported by Escera translates to smaller time scales as well.

We are not the first researchers in the field of music information retrieval (MIR) to recognized that some segments of audio are more important than others. Both Essid [7] and Eronen [5] analyzed the attack separately from the sustained portion of a sound in a "bag-of-frames" approach to instrument classification. In addition, both Jehan [16] and Lyon et al. [20] have utilized audio representations that use non-uniform sampling to emphasize temporally salient regions. Jehan represented songs as time series of audio segments in a number of dimensions that were grouped and sampled on note onsets. We however are interested in comparing individual notes rather than songs. Lyon et al. utilized an auditory model that non-uniformly samples in time using a strobing mechanism to form their audio representation. The main effect of the strobing mechanism is that it creates a stabilized image at times when the audio also sounds stable. While the Lyon approach is an interesting use of auditory models, it does not explicitly assign values of salience to temporal regions.

3. ESTIMATING NOVELTY AS A CUE FOR TEMPORAL SALIENCE

As stated earlier, Escera [6] has shown that novel sounds acquire the involuntary attention of listeners. The time scale Escera was concerned with is much larger than the time scale that we are concerned with. We hypothesize that for our smaller time scale we can use a function that measures short-term local change to estimate novelty *within* audio objects, e.g. single audio notes. We believe that this should also be a cue for the temporal salience of audio which we can use to improve audio similarity measures. We will estimate novelty using two different methods described in the next two subsections: Foote's novelty measure and a power spectral density divergence-based novelty measure.

3.1 Foote's Novelty Measure

The first function we will use to calculate short-term local change is Foote's novelty function [9]. We chose Foote's novelty measure since it is a well-known method that has been cited hundreds of times. This method contains peaks when there is a region with high self-similarity transitioning to a dissimilar region with high self-similarity. The motivation is that at such peaks the dissimilar region will be unexpected, i.e. "novel", given the previous highly self-similar region. This measure is calculated using the following steps:

1. Calculate an audio representation, in our case we used the constant-Q magnitude spectrogram [2]

2. Compute the cosine similarity between each pair of frames and store the results in a self-similarity matrix

3. Correlate a checkerboard kernel with a Gaussian taper [9] down the diagonal of the self-similarity matrix to compute the output. This creates peaks when there is a region with high self-similarity transitioning to a dissimilar region with high self-similarity. The output of this correlation step is the novelty measure.

In our case, the audio was sampled at 44.1kHz and we used a constant-Q magnitude spectrogram [2] with a frame size of 1024 samples, a step size of 256 samples, the Hanning window, and a spacing of 12 bins per octave. We set the width of the checkerboard kernel to 40 frames (250ms).

For more information about this method, see [9].

3.2 Power Spectral Density Divergence Novelty Measure

This measure is of our own creation and captures short-term spectral change by computing a divergence between the power spectral density (PSD) estimates before and after the time frame for which we wish to estimate the novelty. Our idea was to create a measure that captured the amount of new short-term spectral information gained after the point of measurement. The motivation being that peaks in new spectral information likely correspond to unexpected or "novel" events within an audio object.

To represent the spectral content of both noisy and harmonic audio objects, we used averaged periodograms (i.e. Welch's method) to estimate the PSDs [24]. To make these more perceptually relevant, we transformed the PSDs to log-frequency spacing using a constant-Q mapping matrix [4]. We calculated this measurement with the following steps:

1. Calculate the squared magnitude spectrogram using a rectangular window and a DFT size of twice the frame length (i.e. zero pad so as to calculate the Fourier transform of the acyclic autocorrelation function [24]):

$$P_{x_m}(\omega_k) = \left| \sum_{n=0}^{N-1} x_m(n) e^{-j2\pi nk/N} \right|^2 \quad (1)$$

where x_m is the m^{th} rectangular windowed, zero-padded frame of the audio signal, x.

2. For each measurement time frame m:

   (a) Average K frames before and K frames after the measurement time frame m (i.e. compute the smoothed PSD estimate using Welch's method):

$$\hat{Q}_{x_m}(\omega_k) = \frac{1}{K} \sum_{i=m-K}^{m} P_{x_i}(\omega_k) \quad (2)$$

$$\hat{R}_{x_m}(\omega_k) = \frac{1}{K} \sum_{i=m+1}^{m+K+1} P_{x_i}(\omega_k) \quad (3)$$

   (b) Map $\hat{Q}_{x_m}(\omega_k)$ and $\hat{R}_{x_m}(\omega_k)$ from linear frequency spacing to logarithmic frequency (constant-Q) spacing by multiplying both PSDs by a weighting matrix that maps energy in linear-frequency FFT bins to log-frequency bins as defined in [4]

   (c) Normalize the resulting vectors to be distributions and take the Jensen-Shannon divergence [19] between the two.

   (d) The output of the Jensen-Shannon divergence is the output of this novelty measure at time t

In our experiments, we used a frame size of 1024 samples, a FFT size of 2048 points, a step size of 256 samples, a constant-Q spacing of 12 bins per octave, and we set K to 20. We will refer to this measure as "PSD Novelty".

4. EXPERIMENT

We want to estimate the temporal salience of audio to improve audio similarity measures. As mentioned in Section 3, we estimate novelty cues for temporal salience by using two different measures: Foote's novelty measure and a PSD divergence based novelty measure. However, while we believe that these measures may be good approximate cues for temporal salience, this has not been previously established. Therefore, we designed an experiment to answer the following questions:

1. Can either of these novelty measures be used to estimate temporal salience?

2. Can we leverage our estimates for temporal salience to improve an audio similarity measure and therefore improve a machine's ability to predict human similarity ratings of very similar audio objects?

In a pilot study we had subjects try to directly annotate what they perceived as the most salient regions of an audio file. However, subjects found this task very difficult and taxing. Therefore, we used a less taxing but more indirect approach, which is as follows: We had subjects listen to a series of audio clips for which they rated the similarity of pairs of audio clips presented as triplets (i.e. Reference vs. A and Reference vs. B). They listened to three audio clips per trial. First they were played an unmodified reference clip. Then they were presented with two versions of the reference clip that were modified by time compression. One clip was modified in a region of high novelty, and one clip was modified in a region of low novelty. The subjects were then asked to rate how similar each modified audio clip was to the reference audio clip.

Our prediction was that listeners would assign more importance to regions which have high novelty when performing audio similarity judgments, and that they would therefore rate audio clips with modified high novelty regions as less similar to the original audio clip than clips with modified low novelty regions. If this is the case we would be able to affirmatively answer Question 1.

4.1 Stimuli

4.1.1 Controlling for conflated cues in the reference examples

We limited our dataset to artificial sounds so as to avoid conflating any effects with those associated with changes in the physical processes that produce natural sounds. To generate our artificial dataset, we synthesized roughly 1200 audio clips from presets of Native Instruments software synthesizers. The audio clips were six seconds long and were synthesized using MIDI note 60 (middle C). All audio clips were monaural and recorded at a sampling rate of 44.1 kHz and a bit depth of 16.

Recall that loudness and temporal proximity-to-the-onset of an event are cues for temporal salience (See Section 2). Since our goal is to study the effect of novelty, we took steps to control for these other two cues. To control for the effect of loudness, we highly compressed the dynamic range of all of the audio clips, ensuring a rectangular amplitude envelope. To control for the cue related to the temporal proximity-to-the-onset, we added in reversed copies of the audio clips as well. That controls for the temporal proximity-to-onset cue by balancing the stimulus set so that the statistical effect of the temporal onset will wash out.

4.1.2 Determining regions of novelty

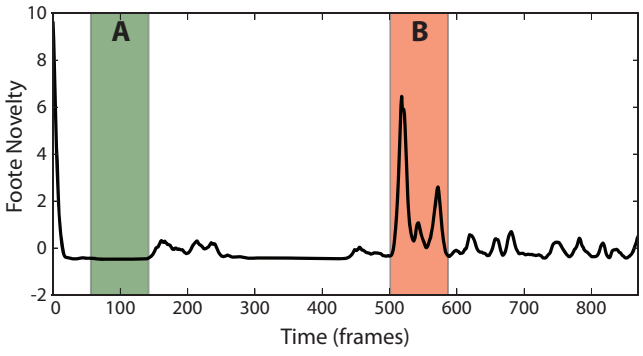

Figure 1: Foote novelty curve. Region A is the 0.5 second window with the lowest novelty, and region B is the 0.5 second window with the highest novelty.

To determine which regions to consider high and low nov-

elty regions, we first computed the Foote novelty function (see Section 3) for each reference audio clip. To choose the regions of highest and lowest novelty, we slid a fixed 0.5 second window across the novelty curve and integrated at each step. The window with the highest integral under the novelty curve was considered the highest novelty region, and the region with the lowest integral was considered the lowest novelty region (see Figure 1).

We used only one novelty function here in order to avoid adding another independent variable. We chose the Foote novelty function due to its accepted use and popularity.

4.1.3 Creating the reference and modified audio clips

We made two copies of each reference audio clip. Using the phase vocoding technique [28], we time compressed the highest novelty region of one copy by a factor of four to create the *high-novelty clip*. To create the *low-novelty clip*, we time compressed the lowest novelty region of the other by a factor of four. This creates an *audio clip group* (the reference clip, the high-novelty clip and the low-novelty clip) of highly-similar clips that we are interested in. We chose time compression as our modification technique because we wanted to minimize changes in novelty to the low novelty region.

From the 1200 audio clip groups, we selected 49 audio clip groups with three criteria in mind: maximize the difference between the integrals of the high and low novelty regions of the reference clip, maximize the diversity of timbre between clip groups, and minimize unintentional phase vocoding artifacts.

As mentioned in Section 4.1.1, we also created time reversed copies of the 49 selected audio clip groups and added them to the dataset, bringing the total number to 98.

4.2 Subject Population

We had 19 subjects in total. All subjects used in the study were adults (18 or older), and the average age of our subjects was 23.9 years. All of our subjects had a at least 6 years of recent, active musical experience on an instrument and/or are familiar with audio processing. The subjects had an average of 11.25 years of musical experience (median: 11).

4.3 Procedures

Subjects were consented one at a time. Subjects were seated in sound isolation booth with a computer that controlled the experiment and recorded their responses. The stimuli were presented binaurally over headphones.

There were 98 unique trials, as well as 5 repeated trials at the beginning for practice, and 19 repeated trials at the end to evaluate each subject's self-consistency. The order of the trials was randomized for each subject. One audio clip group was presented in each trial. Subjects were asked to rate how similar each modified clip was to the reference. Each rating was done using a slider labelled "very dissimilar" and "very similar" on the extrema, and "average level of similarity" in the middle. It took each subject about one hour to complete the experiment.

5. RESULTS

5.1 Consistency

Our ability to use the human similarity ratings for evaluating estimates of temporal salience is dependent on the reliability of the subjects. We therefore looked at the Pearson correlation between the ratings of each subject's test and retest trials (self-consistency). According to this correlation measure, it seemed that subjects had varying difficulty with the task. Self-consistency ranged from 0.047 (random) to 0.81 (highly self-consistent), with a mean (using the Fisher r-to-z transform) of 0.53 and a median of 0.53.

Since we are concerned with how listeners with well trained ears rate similarity, we made the assumption that if subjects have well-trained ears, they will be more self-consistent. We therefore created a subset of subjects for analysis which consisted of all subjects with a high self-consistency (> 0.5). 11 of the 19 subjects were in this subset. The mean self-consistency of this subset was 0.64 (as was the median). From now on we will refer to the full set of subjects as the AS (all subjects) set, and the high self-consistency subset as the CS (consistent subjects) set.

Since all of the human similarity ratings are grouped into trials (one high novelty clip rating and one low novelty clip rating), by comparing the ratings in each trial, we can also view them as the results of a forced-choice experiment, i.e. "which clip is more similar to the reference clip?". This is similar to the *comparison oracle* setup in the machine learning community [11, 18]. When viewing the experiment from this perspective, self-consistency is the fraction of times the same audio clip was chosen as more similar in each subject's test and retest trials. The mean value of this self-consistency measure was 0.74 for the AS set and 0.79 for the CS set.

We also looked at a form of inter-subject agreement on trials by audio clip groups. If subjects had little agreement on particular audio trials, then we can conclude that these trials are confusing. We created a subset of audio clip groups for analysis which consists of the audio clips in the upper two thirds, when ordered by agreement. We will refer to this higher agreement subset as the *high-agreement clips* set, and the full set of audio clips as the *all clips* set (abbreviated HC and AC respectively). Combined with the two subject sets, there are four set combinations, which we will denote as $AS \cap AC$, $CS \cap AC$, $AS \cap HC$, $CS \cap HC$.

5.2 Evaluation of the proposed novelty measures as cues for temporal salience

In this section, we seek to answer Question 1, in which we asked "Can either of these novelty measures be used to estimate temporal salience?"

To answer this, we first standardized the human similarity ratings per subject. We then consider the mean of the pooled ratings between high novelty clips and reference clips, and between low novelty clips and reference clips. The mean human similarity rating for the high-novelty clip is less than that for the low novelty clip in all cases. The consistent subjects judging the high-agreement clips ($CS \cap HC$) showed a mean similarity of 0.48 (SD 0.74) between the reference and the low-novelty clip, and a mean similarity of -0.56 (SD 1.02) between the reference and the high-novelty clip. All subjects judging all clips ($AS \cap AC$) showed a mean similarity of 0.36 (SD 0.84) between the reference and the low-novelty clip, and a mean similarity of -0.38 (SD 1.03) between the reference and the high-novelty clip. All of the paired mean differences are statistically significant ($p < 0.0001$ using the Wilcoxon signed rank test). It is therefore clear that there is at least a relationship between Foote novelty and temporal salience, but how strong is this relationship?

To evaluate the strength of this relationship and that of the PSD novelty measure, we first calculated the novelty time series of each measure for each of the modified audio clips and integrated over the modified regions in each novelty time series. We considered this integral to be the predicted temporal salience of that region. Since we showed in Section 5.1 that the subjects had difficulty making consistent continuous similarity judgements, we again took the *comparison oracle* [11] approach and treated the experiment as a forced-choice experiment, i.e. "which clip is more similar to the reference clip?". The agreement measure shown in this figure is the same as asking what fraction of the time is the following equality true: $(highClipRating \leq lowClipRating) = (highClipNovelty \geq lowClipNovelty)$. We can see from the figure that both the Foote and the PSD novelty measures agree with the human similarity ratings 85% of the time in $CS \cap HC$. From this it seems that the novelty measures are capable of estimating temporal salience in a discrete, forced-choice manner.

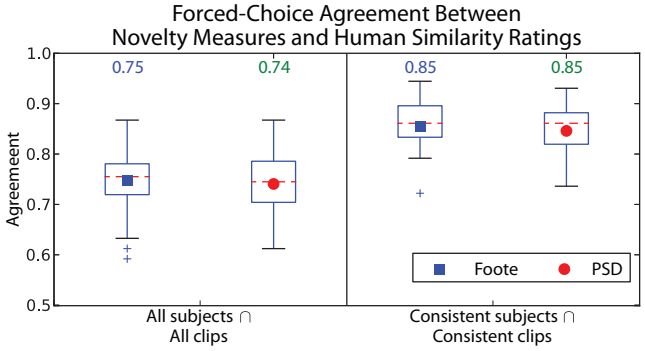

Figure 2: The forced-choice agreement of the novelty measures vs. human similarity ratings.[2]

5.3 Evaluation of distance functions incorporating temporal salience measures

We now evaluate how much the novelty-based temporal salience estimators can improve machine-based audio similarity in the context of highly-similar audio files. To do so, we used two audio distance measures: a "bag-of-frames" measure and a "dynamic time warping" measure. The baseline "bag-of-frames" (BOF) distance function represents each audio clip as a distribution of Mel frequency cepstral coefficients (MFCCs) and compares audio clips by comparing these distributions as described by Aucouturier in [1]. We incorporated the novelty measure into this distance function by using it to weight the probability of MFCCs in the distribution, i.e. we gave more weight to MFCCs that occurred during a high novelty region.

The baseline dynamic time warping (DTW) distance measure was the classic dynamic programming alignment algorithm with an added slope constraint (constrained between 0.5 and 2) as described by Sakoe [23]. The baseline used euclidean distance between MFCC vectors for all path costs. We incorporated the novelty measure into this function by replacing the fixed edit costs (path costs associated with edits) with values of the novelty function.

For both distance functions, we evaluated how incorporating each of the novelty measures compared to the equivalent distance function that does not incorporate novelty. We evaluated the distance measures against the human similarity ratings using forced-choice agreement, i.e. what fraction of the time is the following equality true: $(highClipRating \leq lowClipRating) = (highClipDistance \geq lowClipDistance)$.

As shown in Figure 3, the effects observed by incorporating a novelty function into both BOF and DTW distance functions are very similar when evaluated on forced-choice agreement. Therefore to save space in the remainder of the paper, we will focus only on DTW distance function.

The Foote and PSD novelty measures perform almost identically. Using a DTW distance function that incorporates either the Foote or PSD novelty measure results in mean forced-choice agreement with human similarity choices that are significantly better than that of the baseline ($p < 0.01$ for all sets using the Wilcoxon signed rank test).

Using the PSD-informed DTW distance, we see the following gains in forced-choice agreement over the baseline: 10.29% ($AS \cap AC$), 9.86% ($CS \cap AC$), 10.96% ($AS \cap HC$), 11.69% ($CS \cap HC$). With this, we can affirmatively answer Question 2 – "Can we leverage our estimated cues for temporal salience to improve an audio similarity measure's ability to predict human similarity ratings of very similar audio objects?".

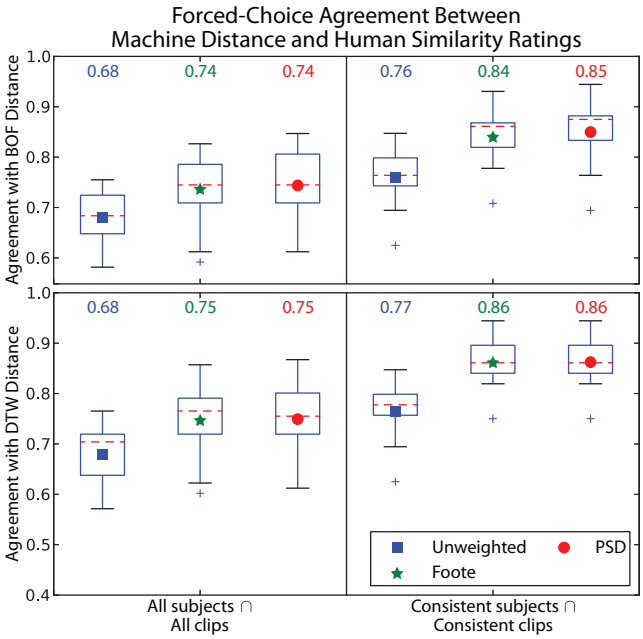

Figure 3: (Top) Forced-choice agreement between bag-of-frames distance and human similarity ratings. (Bottom) Forced-choice agreement between dynamic time warping distance and human similarity ratings.[2]

6. CONCLUSIONS

We proposed two measures to estimate novelty as a cue for temporal salience within audio objects. A human sub-

[2]The boxes extend from the lower to upper quartiles of the data. The dotted lines are the medians, and the symbols (square, star, or circle) in the box are the means. The values at the top of the plot are also the means. $N = 19$ for the AS set and $N = 11$ for the CS set.

ject study where subjects were asked to rate the similarity of very similar audio clips in triples (i.e. "How similar is A to C? How similar is B to C?") shows that most subjects found it difficult to give reliable continuous similarity ratings. The subjects did however seem reliable if we created a forced-choice response from their data, instead answering the question "Which is more similar to C: A or B?"

Figure 2, shows a relationship between salience and both PSD and Foote novelty equally. Incorporating either Foote novelty or PSD novelty into two different audio distance measures showed statistically significant improvements in predicting human similarity ratings on the two-way forced-choice similarity evaluation task. Given a query, a ranking can be returned using such pairwise comparisons in a "tournament" fashion. Therefore incorporating one of these novelty measures could improve audio similarity rankings between highly-similar sound objects.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-0824162 and by National Science Foundation Grant No. IIS-0812314.

8. REFERENCES

[1] J.-J. Aucouturier and F. Pachet. Improving timbre similarity: How high's hte sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[2] J. C. Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[3] M. Caetano, J. Burred, and X. Rodet. Automatic segmentation of the temporal evolution of isolated acoustic musical instrument sounds using spectro-temporal cues. In *Proc. of 13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, 2010.

[4] D. Ellis. Spectrograms: Constant-q (log-frequency) and conventional (linear). http://labrosa.ee.columbia.edu/matlab/sgram, 2004.

[5] A. Eronen. Comparison of features for musical instrument recognition. In *Proc. of IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 19–22, 2001.

[6] C. Escera, K. Alho, I. Winkler, and R. Näätänen. Neural mechanisms of involuntary attention to acoustic novelty and change. *Journal of cognitive neuroscience*, 10(5):590–604, 1998.

[7] S. Essid, P. Leaveau, G. Richard, L. Daudet, and B. David. On the usefulness of differentiated transient/steady-state processing in machine recognition of musical instruments. In *Proc. of 118th AES Convention*, Barcelona, Spain, 2005.

[8] J. Foote. Content-based retrieval of music and audio. In *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, volume 3229, pages 138–147, 1997.

[9] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proc. of IEEE International Conference on Multimedia and Expo*, volume 1, pages 452–455 vol.1, 2000.

[10] B. Glasberg and B. Moore. A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, 50(5):331–342, 2002.

[11] N. Goyal, Y. Lifshits, and H. Schütze. Disorder inequality: a combinatorial approach to nearest neighbor search. In *Proc. of the International Conference on Web search and Web Data Mining*, pages 25–32, Palo Alto, California, USA, 2008. ACM.

[12] J. M. Grey. Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5):1270–1277, 1977.

[13] J. Hajda. A new model for segmenting the envelope of musical signals: The relative salience of steady state versus attack, revisited. In *Proc. of the 101st AES Convention*, 11 1996.

[14] D. Huron. *An Ear for Music*. http://musicog.ohio-state.edu/Music838/course.notes/ear.toc.html, 2002.

[15] P. Iverson and C. L. Krumhansl. Isolating the dynamic attributes of musical timbre. *The Journal of the Acoustical Society of America*, 94(5):2595–2603, 1993.

[16] T. Jehan. *Creating music by listening*. PhD thesis, 2005.

[17] K. Jensen. Envelope model of isolated musical sounds. In *Proc. of the Workshop on Digital Audio Effects (DAFx99)*, Trondheim, Norway, 1999.

[18] A. Karbasi, S. Ioannidis, and L. Massoulié. Content search through comparisons automata, languages and programming. volume 6756 of *Lecture Notes in Computer Science*, pages 601–612. Springer Berlin / Heidelberg, 2011.

[19] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[20] R. Lyon, M. Rehn, S. Bengio, T. Walters, and G. Chechik. Sound retrieval and ranking using sparse auditory representations. *Neural Computation*, 22(9):2390–2416, 2010.

[21] B. Moore, B. Glasberg, and T. Baer. A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240, 1997.

[22] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM, 2003.

[23] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

[24] J. O. Smith. *Spectral Audio Signal Processing*. http://ccrma.stanford.edu/ jos/sasp/, 2012.

[25] J. Thayer, Ralph C. The effect of the attack transient on aural recognition of instrumental timbres. *Psychology of Music*, 2(1):39–52, 1974.

[26] F. Vachon, R. Hughes, and D. Jones. Broken expectations: Violation of expectancies, not novelty, captures auditory attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2011.

[27] D. L. Wessel. Timbre space as a musical control structure. *Computer Music Journal*, 3(2):45–52, 1979.

[28] U. Zölzer and X. Amatriain. *DAFX : digital audio effects*. Wiley, Chichester ; New York, N.Y., 2002.