2007－MUS－72 (7)
2007／10／11

短い楽曲抜粋の中での拍単位の置き換えに関する聴取実験

Sebastian Streich
ヤマハ株式会社
sstreich@beat.yamaha.co.jp

あらまし　拍単位の置き換えに関する聴取実験の結果を報告する．まず，楽曲から 3 拍の長さの抜粋を用意し，その 2 拍目の部分を，別の楽曲から選んだ 1 拍に置き換えたものを作った．被験者は，この抜粋を聴き，置き換えの適切さを，印象に基づいて 3 段階で評価した．置き換え用の 1 拍を選ぶのに，3 種類の方法を用いた．1 つめは音の開始点を基準とする方法，2 つめは一定数のフレームをグループとして扱う方法，3 つめはランダムに選択する方法である．被験者評価の全体平均は，中程度の満足度を超えることはなかったが，統計的には有意な差異が見出された．音の開始点を基準とする方法が，もっとも高い被験者評価を得た．ランダムに選択する方法はもっとも低い評価となった．

A Listening Experiment on Beat Substitutions in Short Musical Audio Excerpts

Sebastian Streich
YAMAHA Corporation, Hamamatsu, Japan

Abstract: We report the results of a listening experiment on beat-based substitution. First, music excerpts that consisted of three consecutive beats were prepared. The central beats of those were then substituted by a beat from a different music track. Subjects listened to the excerpts and rated the suitability of these substitutions on a three point scale. Three methods for substitution selection are compared: an onset-based method, a fixed-length frame grouping, and a random selection. While overall average ratings didn't exceed intermediate satisfaction, we found statistically significant differences. The onset-based method clearly received the highest ratings, while random selection received the lowest.

1. Introduction

In this article we are addressing a line of research where the properties of very short music excerpts are considered and compared. Different authors have coined different terms for this domain. Among the most used are: *Concatenative Sound Synthesis* [01], *Audio Mosaicing* [02], or simply *Musaicing* [03]. The purpose here is to generate new music by recombining short fragments of existing sound or music recordings (roughly in the range of a couple of hundred ms). This generation can be fully automatic or allow for some type of interactivity of the user/composer with the system, but in any case it is guided by computed descriptions that are attached to each fragment. Most approaches focus on artistic or experimental aspects, which can't be evaluated with standard tools of science. Therefore the different computational approaches for comparing and selecting the fragments are normally chosen based on subjective preference or experience.

Here, we want to consider a very specific type of setting for Audio Mosaicing: the case where the fragments correspond to the beats of music audio recordings. For our experiment we narrow the specifications down further by focusing on a fully automatic approach where a target beat is being substituted by a fragment selected from a database holding a wide variety of music recordings. The system needs to identify fragments that are similar to the target beat and make an acceptable substitution for a human listener. We conducted a listening test where subjects were asked to express their acceptance for each one of three different substitutions on a three-point scale. In the following sections we will first explain the chosen methods for beat selection. Afterwards we are going to describe the experimental setup and the results obtained from the statistical analysis of the collected data. Finally, we present our conclusions and ideas for future work.

2. Beat Selection Methods

The main aspect we wanted to examine with this experiment consisted in a comparison of different representations of the beat properties. We wanted to test their influence on the human listener's impression on the resulting substitutions. Many of the descriptors that are commonly used

in audio mosaicing are genuinely computed on single, overlapping STFT frames of 10–100 ms duration. That means a beat of for example 750 ms might easily comprise more than 100 different values of a single, scalar descriptor. Keeping all of those, however, is not practical for several reasons. First, it would require a lot of memory and processing time to search for a match in the database. Secondly, it is desirable to achieve a more generalized representation, since we are interested in a similar, not in an identical match.

In contrast to the case of similarities of full songs, representations considering only plain statistical properties on the beat level are missing an important aspect for the chosen setup: the timing of the musical events within the beats. We tested two methods to overcome this problem by considering sub-segments within each beat. Both methods are starting exactly from the same frame-level features. For our experiment we restricted the selection to two elementary features in the style of Jehan's approach [04]. They were computed for each STFT frame of 1024 samples with 50% overlap.

The timbre and loudness properties were both covered by the bark band energies, a 25-dimensional vector with the accumulated energies of the STFT bins falling into the same bark band (see Zwicker [05] for a definition of bark bands). As a preprocessing step we applied a filter which mimics roughly the frequency response of the human auditory system. The accumulated energy values were converted to log-scale.

The pitch properties were reflected in the harmonic pitch-class profile (HPCP) developed by Gomez [06]. This 36-dimensional vector contains basically the accumulated energies for each pitch class of the well-tempered scale mapped into a single octave at a resolution of 1/3 semitone. Each vector has its maximum normalized to one. For details on the computation please refer to [06].

2.1 Method A: Onset-based Grouping

For this method we first utilize an onset detection algorithm (provided from a research collaborator) in order to obtain the locations of sound onsets in the music signal. State-of-the-art, fully automatic methods achieve around 70% of correct detections (average F-measure, see [07]). However, for testing the approach under realistic conditions and also for practical reasons we did not perform manual annotation or editing of the onset locations.

In a second step we then unified the beat and onset locations. This was done by applying heuristic criteria on the data with two objectives: avoiding fragmentation into too short pieces, and focusing on major changes in the music audio signal. Onsets refer to the beginning of a new sound event, which in general causes an increase in acoustic energy in the audible spectrum. The onset detection algorithm located the onsets at the energy peaks, while we wanted to consider the peak in energy increase. Therefore we computed the first order difference function for each dimension of the bark band vectors. We accumulated only the positive values (increasing energy) across all dimensions for each frame transition. This accumulated energy increase was then used to adjust the exact onset positions within a range of four frames (~46 ms) to the front and one (~12 ms) to the back. In case more than one onset was located within a range of seven frames (~81 ms) only the one with the highest energy increase was kept. Then we performed the actual unification of the beat locations with the onset locations. If an adjusted onset was found within the seven frame range (~81 ms) of a beat, we adjusted the beat location to coincide with the onset. Otherwise the beat location remained unchanged.

With the beat and onset locations prepared we could then arrange the storage of the feature data for each beat. As for the bark band vectors we decided to simply store the data of the first 10 frames of each onset. The rationale behind this approach was to preserve the attack period as accurately as possible, because it contains the most important information about the timbre of the starting sound event (see e.g. Grey [08]). A second aspect is that in order to adjust the length of the substitute beat to that of the target beat we simply chop off the end of each sub-segment accordingly.

For the pitch related feature we followed a slightly different approach. In the first place we wanted to be able to distinguish between sub-segments that contained clear and stable pitches, and sub-segments with highly varying pitch or dominant percussive sounds. To achieve that we introduced a "tonalness" measure T_k that is computed for each HPCP vector $hpcp_k$ according to the following formula:

$$T_k = 50 - \left(\sum_{i=1}^{36} hpcp_k(i) \right)^2 \quad (2.1).$$

The measure takes advantage of the normalization of the HPCP vectors. Two extreme examples might illustrate this. If we consider a

frame that contains exclusively a single sinusoid, we would find the entire spectral energy concentrated within a single pitch class. After normalization we would obtain a vector like (0 0 0 1 0 0 0 0 ...). This vector has a "tonalness" value of 49. On the other hand a frame that contains a noise-like sound with its energy spread across the entire spectrum would be found to have a rather flat distribution on the different pitch classes. After normalization it might look similar to (0.93 0.89 0.96 1 0.92 ...). The corresponding "tonalness" is far below zero. For our experiment we fixed the lower limit of T_k to zero.

For the database we computed a weighted sum of the corresponding HPCP vectors of each sub-segment. The weight for each vector was determined by two components: its perceptual energy (obtained from the bark band vectors) and by its value T_k . The first component was normalized to add up to one for each sub-segment, while the second was left unchanged.

It is important to note that for the retrieval of substitute beats with method A we searched across beat boundaries. Since we preserved the information about the original order of the beats (see figure 2.1.), we could simply treat the target beat as a sequence of sub-segments and look for a substitute sequence without considering the beat locations.

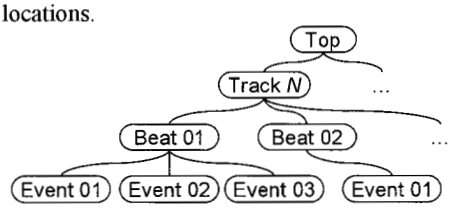

Figure 2.1: Database structure for method A.

We followed a stepwise procedure to find the ideal substitute. First, we applied a minimum length criterion to make sure each retrieved sub-segment was at least as long as its corresponding target segment. Then we further reduced the search space by comparing the "tonalness" values. Since the "tonalness" was applied as a weight to each HPCP vector without normalization we can simply use the maximum value of the accumulated HPCP vectors as an indicator for the stability and clarity of the pitched content in each sub-segment. The criterion we applied consisted in a tolerance margin around the "tonalness" level of each target sub-segment. We allowed a margin of ±15. Figure 2.2 illustrates exemplary the effects of the two pre-selection steps for method A (darker bars). The number of candidates before pre-selection was around 23.000 in each case.

For the remaining candidates we computed the Euclidean distance between the bark band vectors to the target. This distance was multiplied with the Cosine distance between the accumulated HPCP vectors. Finally, we multiplied the distance values for each sub-segment in the candidate sequence to achieve a final distance score. We then picked the candidate sequence with the minimum distance as the substitute. Each sub-segment of the selected substitute was chopped in order to obtain a sequence of the same length as the target sequence.

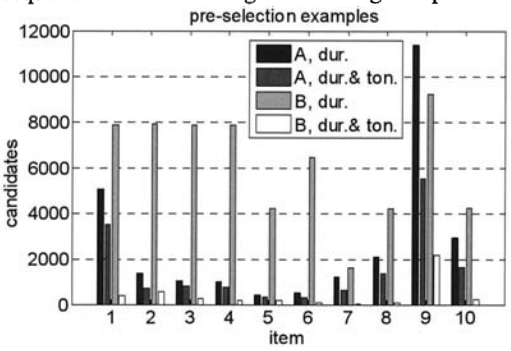

Figure 2.2: Number of candidates after the two pre-selection steps of method A and B for the first ten items in the test set.

2.2 Method B: Fixed Length Grouping

This method operates on a lower abstraction level than method A. For a fair comparison we used the modified beat boundaries that we obtained from method A, but otherwise the onsets were not considered at all. Instead we performed a decimation of the feature data by simply grouping together every five consecutive frames (~60 ms). For the HPCP vectors the database was generated in the same manner as with method A. We followed a similar approach for the bark band vectors this time, except that the "tonalness" was not considered as a weight here.

With method B we queried directly on the beat level. Again, a stepwise procedure was applied. First, only beats with at least the same length as the target beat were considered potential candidates. Other than with method A we adjusted the length of each candidate by simply chopping off the part exceeding the target beat. Secondly, the same "tonalness" criterion as with method A was applied. The "tonalness" of all groups of the potential substitute had to lie within the tolerance margin of their corresponding target group. The effects of the two steps for method B are illustrated in figure 2.2 above (lighter bars). The number of candidates before pre-selection was always 9490. To compute the final distance score

we then followed the same procedure as described for method A.

2.3 Method C: Random Selection

This selection method was included as a baseline reference in order to get a better idea of the absolute performance of the other two methods. Substitute beats were chosen completely at random with only two restrictions: their length had to be greater or equal to the target beat, and they had to be from a different track than the target beat.

3. Experimental Setup

3.1 Selection of the Audio Data

The music database for this experiment comprised a total of 171 30sec-long excerpts from distinct music tracks adding up to a total of 9490 beats. Tracks originated from a wide variety of genres and styles containing Electronic Dance, HipHop, Rock, Pop, J-Pop, Acid Jazz, Blues, Funk, Dub, Punk, Metal, Classical, Oldies, Latin, and World. The audio data is of commercial CD-quality and had been converted to 44.1 kHz, mono format. The beats had been manually annotated for the excerpts by two professional musicians.

In a first stage we selected two beats from each track in the database at random. We then used these 342 beats as targets for the three different methods. Since we wanted to obtain ratings by each subject on the entire test set, we had to significantly reduce the number of items. To do so we first removed those items for which at least two of the methods had found an identical substitution. The rationale behind this step was to focus on the differences in performance between the three approaches. This left us still with 305 items in the data set. For further reduction we selected the 28 best matches in terms of the similarity measures for each of the methods A and B. By chance this selection gave us a three third mixture consisting of 14 tracks appearing in both lists, 14 tracks appearing in list A but not in B, and 14 tracks appearing in list B but not in A.

3.2 GUI of the Experiment

We designed a simple GUI in Tcl/Tk using the SNACK package to handle audio playback. The GUI is shown in figure 3.1 below. Subjects were enabled to listen as many times as they liked to the original and the modified three-beat samples. They could rate each of the modified samples individually on a three-point scale with the levels "Bad!", "OK.", and "Good!". A neutral category named "Can't tell." was additionally provided. Once the subject clicked the submit button the ratings were stored in a text file and could not be modified anymore.

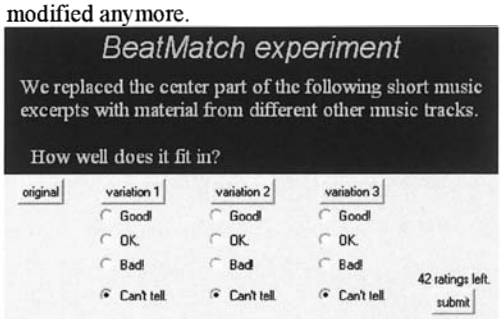

Figure 3.1: Snapshot of the GUI for the listening experiment.

The sequence in which the items were presented was randomized for each subject. Also the assignment of the methods to the buttons was randomized for each single rating. Subjects had no knowledge about the beat selection methods applied in the experiment.

4. Results

A total of 20 subjects participated in the experiment. Each of them completed the entire test unsupervised at their own computer and sent back the results by email. We analyzed the obtained data with statistical methods in several ways.

4.1 Average Ratings per Method

We compared the average ratings for each of the methods across all items and subjects. As can be seen from figure 4.1 method A received the best ratings, followed by method B, with method C being the worst of the three. All differences are statistically significant (with $p>0.99$). Neutral ratings (category "Can't tell") were excluded from the averaging. They accounted for less than 3.7% of the ratings with either method.

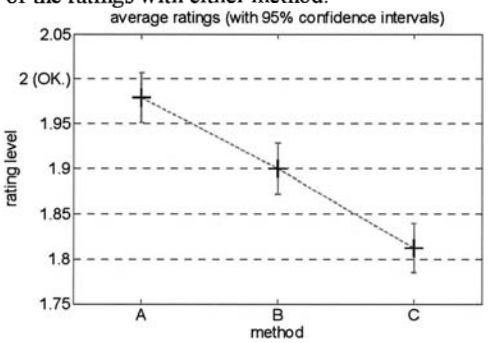

Figure 4.1: Overall average ratings (1=Bad, 2=OK, 3=Good).

We can also observe that overall only an intermediate level of satisfaction was achieved despite the selected items were already taken from the best 10% of matches in the database according to the applied distance measures.

4.2 Rater Agreement

We computed raw rater agreement indices for the different satisfaction levels for each method. The purpose of this analysis was to see how subjective the ratings are. Raw rater agreement indices reflect the proportion of actual agreements among raters compared to the maximum possible number of agreements.

In figure 4.2 we provide the total agreement indices as well as the specific ones for each rating category as the broad colored bars. The white rectangle within each bar corresponds to the 95% confidence intervals for a randomized distribution with the same base rates. The emerging pattern is that subjects tend to agree significantly beyond chance level on the two extreme categories. In contrast, the agreement for the neutral and the intermediate categories hardly reaches statistical significance and could be attributed mostly to chance.

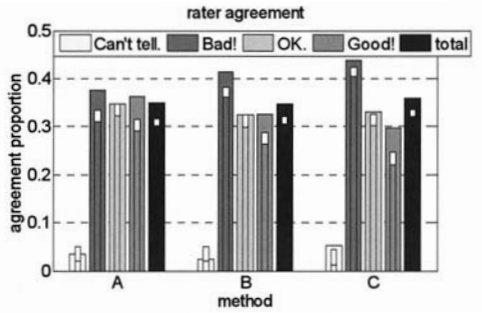

Figure 4.2: Raw rater agreement indices with 95% confidence intervals (white rectangles).

4.3 Pearson Correlations

As a third type of analysis we computed several correlation coefficients with other variables. At first we tested whether the rating levels were correlated with the rating order, which was not the case.

The next correlations were based on the average ratings for each item and method across all subjects. Here we first tested for correlations between the ratings for the different methods. All pairs turned out to be highly correlated with 32-55% of the variance being explainable by the rating for the other method (see table 4.1).

method \ method	B	C
A	*.57*	*.62*
B	-	*.74*

Table 4.1: Correlation coefficients for average ratings between the different methods.

This could be caused by two things. It could be that for certain types of music the beat replacement in general works better or worse than for other types. So no matter which selection method is used, the results will always be rather good or rather bad. On the other hand, since we always presented the substitutions for all three methods in parallel, it could simply be the case that subjects had the tendency to adjust their ratings close to each other avoiding big discrepancies.

Another interesting observation was that for each of the three methods the ratings showed negative correlations with the target beat duration. So it was in general perceived more disturbing when a beat in a slow musical piece was replaced than in the case of a fast piece. This correlation was the strongest in the case of random selection (r=-.43). However, there was no significant correlation with the number of detected onsets in the target beat, neither with the distance score of the selected substitutions. The latter is particularly noteworthy, because it shows a deficit in the distance computation methods to reflect human judgments on the finer scale.

As a final test we tried to investigate the effect of the pre-selection steps during the retrieval procedure. To do so we correlated the average ratings per item with the number of potential substitutes left after each pre-selection step. The results are shown in table 4.2, statistically significant values are slanted.

Pre-selection \ method	A	B	C
1. duration	.21	*.34*	.19
2. "tonalness"	.05	.09	*.42*

Table 4.2: Correlation coefficients for average ratings with number of candidates after pre-selection steps.

Basically all results show no significant correlation with two exceptions. The duration pre-selection seems to have a bigger effect for method B than in the case of method A, despite it is much less restrictive (see figure 2.2). It is notable that with the further reduction by applying the "tonalness" criterion the number of candidates becomes completely uncorrelated again for method B, although this step means a massive reduction in most cases. For method C the "tonalness" criterion was not applied. The clear

positive correlation can be explained by considering that it is more probable to pick up an acceptable replacement by chance when there are many suitable candidates in the set.

5. Conclusions

We reported the results of a listening test comparing three different methods for automatic beat substitution in short musical audio excerpts. Our results revealed that an approach considering onset locations – even though error prone – might reach better acceptable results in automatic beat substitution. Our observations from the correlation tests indicate that a criterion to pre-select substitution candidates based on their "tonalness" helps in making the search much more efficient. Overall the feature representation and retrieval methods still need to be improved as they currently only achieve intermediate acceptance.

6. Future Work

For audio mosaicing applications like the one reported here the size of the database is critical. Even short musical excerpts contain a big variety of distinct properties that are impossible to match if there simply is no adequate item in the database. Future work might consider allowing certain modifications of particular properties in order to provide better coverage of the parameter space with a smaller database. Prominent examples would be adjustments of loudness or pitch. This requires of course a distance metric which takes these adjustments into account during the search.

7. Acknowledgements

The author would like to thank his colleagues who spent their time participating in this test. Special thanks go to Mr. Takuya Fujishima and to Mr. Keita Arimoto for their assistance in preparation and organization of this experiment.

8. References

[01] D. Schwarz. The caterpillar system for data-driven concatenative sound synthesis. In Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03), September 2003.

[02] A. Lazier and P. Cook. Mosievius: Feature driven interactive audio mosaicing. Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03), September 2003.

[03] A. Zils and F. Pachet. Musical Mosaicing. In Proceedings of the 4th International Conference on Digital Audio Effects (DAFx-01), December 2001.

[04] T. Jehan. Creating Music by Listening. PhD thesis, Massachusetts Institute of Technology, September 2005.

[05] E. Zwicker and H. Fastl. Psychoacoustics: Facts and Models, 2nd ed., Springer Verlag, Berlin 1999.

[06] E. Gómez. Tonal description of polyphonic audio for music content processing. INFORMS Journal on Computing, Special Cluster on Computation in Music, vol. 18:3, 2006.

[07] MIREX wiki-page. Audio Onset Detection Results, 2006. http://www.music-ir.org/mirex2006/

[08] J. Grey. Timbre discrimination in musical patterns. Journal of the Acoustical Society of America, vol. 64, pp. 467-472, 1978.