EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Method for creating location-specific audio textures

Toni Heittola[1*], Annamaria Mesaros[1,2], Dani Korpi[1], Antti Eronen[3] and Tuomas Virtanen[1]

Abstract

An approach is proposed for creating location-specific audio textures for virtual location-exploration services. The presented approach creates audio textures by processing a small amount of audio recorded at a given location, providing a cost-effective way to produce a versatile audio signal that characterizes the location. The resulting texture is non-repetitive and conserves the location-specific characteristics of the audio scene, without the need of collecting large amount of audio from each location. The method consists of two stages: analysis and synthesis. In the analysis stage, the source audio recording is segmented into homogeneous segments. In the synthesis stage, the audio texture is created by randomly drawing segments from the source audio so that the consecutive segments will have timbral similarity near the segment boundaries. Results obtained in listening experiments show that there is no statistically significant difference in the audio quality or location-specificity of audio when the created audio textures are compared to excerpts of the original recordings. Therefore, the proposed audio textures could be utilized in virtual location-exploration services. Examples of source signals and audio textures created from them are available at www.cs.tut.fi/~heittolt/audiotexture.

1 Introduction

Virtual location-exploration services, such as Here Maps 3D [1], Google Street View [2], and Microsoft Streetside [3], provide ways to virtually explore different locations on the globe. In these services, a user is able to see a 360-degree view at various locations. The image material is typically collected using video cameras mounted on cars which drive around the streets of a city. Using an audio ambiance in conjunction with visual information would provide additional information and give the service a more 'real' feeling. In addition, audio ambiance would provide a rich source of information characterizing a location. A location could be, e.g., a busy street corner in New York or a quiet back alley in Tokyo. Users of such an exploration service may be interested in how crowded a certain location is at a certain time of the day, whether there is a lot of traffic noise on the adjacent street when choosing a hotel or in which parks around the potential hotels one can hear birds singing.

Currently, all the available location-exploration services provide only visual description of the location, and no location-specific audio ambiances are used. One of the reasons may be the cost of collecting a comprehensive audio database: to record audio for such a service, one would need to stay at each location for a certain period of time in order to collect enough audio data for non-repeating audio ambiance. Current data collection methods involve driving by with a car while recording images and other type of information. Such method is not suitable for collecting audio material, as the sound produced by the car would interfere with the audio content. As crowd-sourced data is a feasible solution to data collection, it is attractive to be able to create location-specific audio ambiances in a cost-effective way, using a small amount of audio. In addition to the virtual location-exploration services, location-specific audio ambiances could be used in applications such as computer games to enhance the user experience.

Studies show that presenting visual and audio information together enhances perception in comparison to presenting only visual information [4,5]. Adding audio ambiance for the specific location, even having different ambiances depending on the time of the day, would

*Correspondence: toni.heittola@tut.fi
[1]Department of Signal Processing, Tampere University of Technology, P.O. Box 553, Tampere 33101, Finland
Full list of author information is available at the end of the article

enhance the user experience in the virtual location-exploration services. A few works that combine visual and audio representations for a location suggest using abstract or natural sounds for representing additional information [6,7] or using characteristic audio sequence for mapping environmental noise into city models [8]. However, none of these prior works use specific audio ambiances recorded at the location of interest.

According to listening tests, humans are able to discriminate everyday audio ambiances [9,10]. The listening test conducted in [9] showed a recognition performance of 69% with 24 everyday environments such as street, nature, restaurant, or cafeteria. In [10], same type of test was conducted with more descriptive audio categories. Some of the categories included in the test were representing different situations within the same environment, such as nature during daytime or during nighttime and street with traffic or with police car. The experiment reports a recognition performance of 82% with 15 categories.

A more detailed listening test, presented in [11], shows that humans are able to discriminate even specific locations based on the audio, such as different streets of the same urban area. This test was conducted as a preliminary study for developing location-based systems augmented with audio ambiances. The results indicate that audio ambiances from similar urban areas of the same city (e.g., different streets, different parks) differ enough to make them recognizable. Thus, location-specific audio ambiances should use audio recorded at the location of interest.

Our work proposes a method to create a new arbitrary long and representative audio ambiance for a location by using a small amount of audio recorded at the exact location. The method aims to retain the location-specific characteristics of the auditory scene while providing a new and versatile audio signal to represent the location. The steps involved in the method are the segmentation of the source audio signal into homogeneous segments, finding a new segment sequence with smooth timbral transitions, and creation of a new signal by concatenating segments in such a way that the segment boundaries do not create perceivable cuts. This new generated signal is an audio texture. The audio texture is constructed from the segments of the source audio signal by repeating them from time to time but never in their original order, resulting in a continuous, infinitely varying stream of audio. Despite the repeated segments, the audio texture should have a non-repetitive feel to the listener. In order to create location-specific audio texture, the source signal has to be recorded at the given location.

Previous studies for audio texture creation have been mostly concentrating on producing textures which are perceived as homogeneous in an interval of few seconds [12-15]. In these studies, textures are composed of small sound elements, which are occurring according to a specified higher level pattern. Typical target sounds for this type of audio textures are background sounds, e.g., rain, wind, and crowds. We use the term audio texture in our approach even though our goal is to synthesize a signal to represent the whole acoustic scene, not only background sounds. Our method aims to produce more heterogeneous audio textures than the ones typically targeted with audio texture synthesis methods, better suited for realistic auditory scenes. Therefore, longer sound elements and longer pattern intervals than in typical audio texture approaches are considered in our work. The reason for this is to keep homogeneous segments of the auditory scene intact (i.e., segments containing similar sound events) and produce a versatile but realistic representation of the source auditory scene.

Previous work in automatic generation of soundscapes used isolated sound event samples together with audio textures as components to create auditory scenes [16]. Our method avoids explicit usage of sound event samples when creating a synthesized representation of an auditory scene, as it is complicated to automatically find samples with unique content and acoustic properties that match a target scene.

The paper is organized as follows. Section 2 discusses related previous work, and Section 3 presents a detailed description of the audio texture creation method. Section 4 presents the details of a listening test created for the evaluation of the obtained audio textures, and Section 5 presents and discusses the results of the listening test. Finally, Section 6 presents conclusions and ideas for future work.

2 Related work

Previous work on synthesis of audio textures has concentrated on producing sounds with temporal homogeneity [12-14,16]. On the broader scope, the idea of audio textures is shared in many other domains, such as music textures [17] and musical mosaicing [18].

There are many application areas for audio texture synthesis in multimedia creation and games, and a wide range of different approaches have been proposed for it: noise filtering [19], physical modeling [13], wavelets [14], granular synthesis [12], corpus-based concatenative synthesis [16]. A comprehensive review of state of the art in audio texture synthesis approaches can be found in [15]. The method most related to our work, granular synthesis, uses small segments of an original signal, and some statistical model to synthesize the texture using those segments. The granular synthesis approach is extended in corpus-based concatenative synthesis with a large database containing segments from source audio which are organized according to audio descriptors (e.g., pitch and timbre) [20]. In the synthesis stage, the segments are retrieved from

the database according to the specified descriptors and concatenated to form a texture.

Audio texture synthesis method for signals with simple structures is proposed in [12] for background sound creation and audio restoration. The main objective of the work is to produce a continuous and perceptually similar texture to the source signal. Segmentation of original audio is done using novelty detection method applied on the similarity matrix; sequencing of segments is based on transition probabilities.

Soundscape synthesis combining audio textural elements and sound object elements for virtual environments is proposed in [16]. The synthesis elements were retrieved from user-contributed unstructured audio corpus, categorized into these two layers, sound objects being the sounds meant to draw attention and textures being the sounds that form the ambiance. For obtaining elements for textures, segmentation is based on Bayesian information criterion with a minimum segment length separately chosen to three categories (natural, human, mechanical). The texture was created using concatenative synthesis, and sound objects were mixed at the final stage to form the soundscape. A subjective evaluation of the immersive properties of the synthesized soundscapes in virtual environments showed soundscapes to be acceptable in comparison to the original recording.

Music textures are a special case of audio textures, where the underlying musical structure is utilized in the segmentation process. The work in [17] describes various ways to create music textures, by using an onset detection method for selecting segments of audio. The segments between two onsets are the units to be rearranged. The work demonstrates recreating music pieces by selecting similar segments from other pieces and concatenating the segments. Applications include, e.g., creating audio textures, music remixing, and replacing corrupted segments in an encoded audio file with synthesized ones. Another example on creating new musical audio signals is the musical mosaicing presented in [18]. The input of this method is a sequence of higher level constraints, for example pitch. Based on this sequence, samples are retrieved from a large database and joined to create the mosaic.

The main difference between methods for creating music textures and methods which operate on more general signals like everyday audio ambiances is that music rearranging methods can utilize the metrical structure in selecting the segment positions. For example, [17] applies a similarity measure based on the metrical location of a segment. Everyday audio ambiances contain a lot of overlapping sound events, and usually, there is not a clear structure between events which could be utilized in the audio texture creation. In this case, the methods operate on more local constraints trying to ensure, for example, that there are no abrupt changes in the timbre of two adjacent segments.

The concatenation of different segments in creation of an audio texture can in fact be considered similar to the concatenative speech synthesis [21], where units representing phonemes are concatenated in a desired order. Unit selection rules are employed so that the best chain of candidate units is chosen from the database to obtain smooth and natural transitions from one to another. Unit selection is done based on acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighboring phones, with as little as possible processing at the point of concatenation [21]. The concatenation of music segments in [17] does not deal with inconsistencies at segment boundaries. The boundaries are at onsets and this tends to mask the cuts.

There are multiple services available to collect geotagged sound samples, as part of acoustic ecology and soundscape evaluation. The Sound Around You [22] Project [23] collected audio data uploaded by participants using mobile devices. For each uploaded recording, the participants subjectively rated the soundscape in terms of, e.g, quality, pleasantness, and tranquility. The scope of the project was to study the perception of soundscapes and their link to human activities. A collaborative artistic project called the Urban Remix project also collected sounds recorded using mobile devices along with geotagged information [24]. These web-based services allow users to open a map view and listen to the sounds recorded at various locations. In Urban Remix, the user can also create new soundscapes by drawing a path on the map, which will create a mix of the sounds encountered along this path. Practically, these services offer the possibility of listening to the samples, usually of short duration, that other users recorded at specific locations.

Audio synthesis techniques would provide the possibility of creating longer and more versatile audio content to characterize a location. A method for generating location-specific audio is presented in [25] as an environmental sound synthesis for location-based sound search system. Their method uses spatial audio mixing based on geographical features and dynamic situations, requiring additional knowledge about the location (e.g., listener position, position of obstacles such as buildings). Our work has similar goal of creating location-specific audio, and using audio textures synthesis based on short original recording from a location avoids the collection of any additional location-specific information.

3 Audio texture creation

By an audio texture, we mean a new, unique, audio signal created based on a source audio signal acquired from a certain location. The texture should provide an audio representation of the location from which the source signal

comes from by having the same generic acoustic properties and types of sound events that are characteristic to the location.

The proposed audio texture creation method consists of two stages: analysis of the source audio signal and synthesis of the audio texture. The main steps are illustrated in Figure 1. The analysis stage performs clustering and segmentation of the source audio in an unsupervised manner. The goal of the clustering analysis is to automatically find homogeneous audio segments from the source signal. Ideally, the segments would be representative of individual sound events from the auditory scene. In the synthesis stage, an audio texture is constructed by shuffling and concatenating these segments. The shuffling will be done in a way that takes into account the timbral continuity at the segment boundaries.

3.1 Similarity-based segmentation

The segmentation is performed in an unsupervised way by using the self-similarity of the source audio. The source audio is split into non-overlapping 4-s chunks. This means that the system is looking for homogeneous regions that are at least 4 s long. Mel-frequency cepstral coefficients (MFCCs) calculated in short frames are used to represent the rough spectral shape inside the chunks. The MFCCs have been successfully applied in speech and speaker recognition, various music signal classification tasks (artist, genre, song structure), environmental audio classification, audio event detection, and audio context classification [9,26]. For the proposed system, the 16 lower order static MFCCs and their delta coefficients are extracted in 20-ms frames with 50% overlapping. The feature vector also includes the zeroth order MFCC coefficient, which corresponds to the energy of the frame. This is done so that the segmentation process would react also to changes in signal energy.

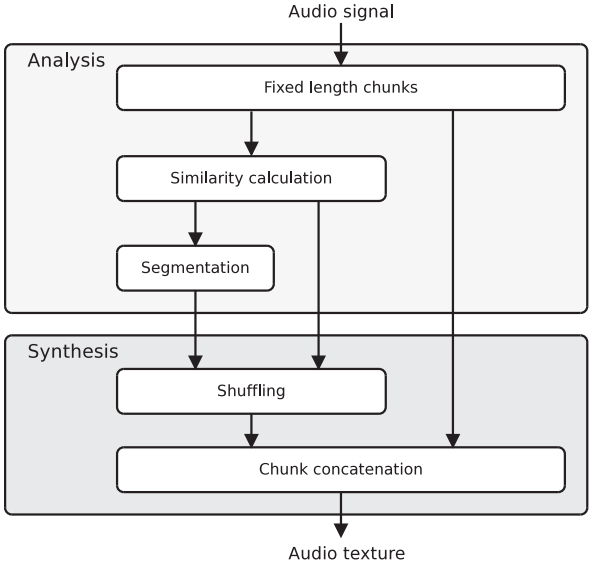

Figure 1 Main steps of the audio texture creation application.

Gaussian mixture models (GMMs) are trained for each chunk. Since the chunks have varying complexity, the agglomerative EM algorithm [27] is used to train GMM with the best amount of Gaussian components individually for each chunk. The used training algorithm is based on an information theoretic criterion for choosing the number of components. In this work, the maximum possible number of Gaussians is fixed to eight.

A distance matrix is formed by calculating the distance from each of the chunk-based GMMs to every other. The smaller the distance, the more similar the chunks are. Distances between two distributions, GMMs, are calculated using the empirical symmetric Kullback-Leibler divergence [28]. The distances between all the GMMs are collected into a matrix, as a representation related to the similarity between the 4-s chunks. Each row is a feature vector representative of the dissimilarity of one chunk to every other. The matrix reveals patterns in the source audio, and the acoustically homogeneous segments will appear as rectangular regions on the diagonal of the matrix [12,29].

Various methods exist for extracting the segment boundaries from a similarity matrix or a distance matrix. A common method, used for example in song structure segmentation, is to calculate a novelty curve, and its maxima are considered segment boundaries [12,29]. We will use a method based on clustering of the matrix rows [30]. The reasoning behind this is that a row of this matrix describes how close one chunk is to every other chunk, and this can be considered as a feature vector for that specific chunk.

A k-means algorithm is used to cluster the rows [30]. In this process, similar chunks of the source audio are clustered together. The number of clusters should ideally be related to the expected complexity of the auditory scene. Since we do not have reliable ways to estimate scene complexity, in this study, the number of clusters is fixed to 20. A chunk merging step assigns to the same segment the consecutive chunks (in time) that belong to the same cluster.

A simplified example of the distance matrix and clustering is shown in Figure 2. In this case, chunks 1 and 2 are similar because the calculated distance between them is small. They are both similar with chunk 4 since the calculated distances $d(1,4)$ and $d(2,4)$ are also small. During the clustering, these chunks will be assigned to the same cluster. At chunk merging, chunks 1 and 2 will be merged to form a segment of length 8 s and this will be one unit in the synthesis stage.

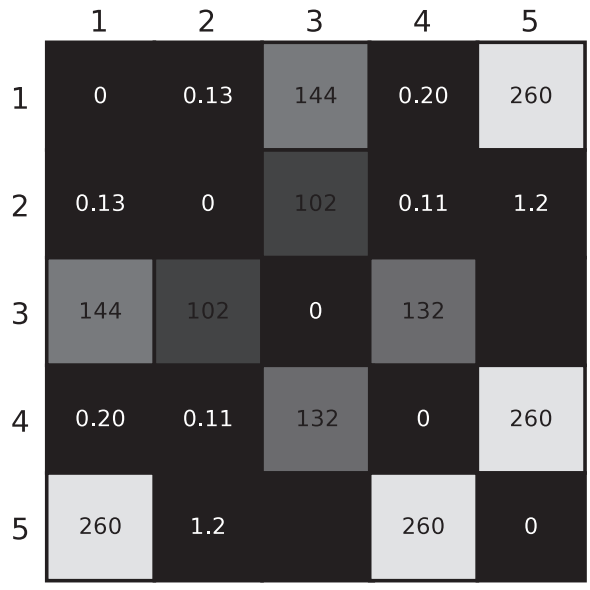

Figure 2 Exaggerated example of distance matrix between chunks.

Figure 3 presents a segmentation example. The upper panel shows the distance matrix, in which the darker color corresponds to smaller distance between chunks. The lower panel shows the outcome of the segmentation process. Each color in the lower panel corresponds to a specific cluster, and adjacent chunks belonging to the same cluster form a segment. The resulting segments are indicated also in the distance matrix by the rectangles on the diagonal. Their lengths are multiples of chunk length.

The outcome of these steps is a collection of segments of audio that are timbrally homogeneous, spanning all the source audio data. These segments will be used to create the texture by concatenating them as it will be explained in the following section.

3.2 Shuffling and concatenation

The synthesis stage of the system consists of a shuffling controlled by the similarity of the segments and their concatenation for obtaining the audio texture. As illustrated in Figure 1, the synthesis stage uses the segments obtained after the analysis stage and the similarity information that was also calculated during analysis.

A straightforward method to obtain a new signal based on segments is simply randomly shuffling them. However, this would most likely produce audible artifacts on the segment boundaries. In order to avoid the artifacts, we devise segment selection rules that will take into account timbre continuity at the boundaries of consecutive segments. We guide the segment selection using the similarity between the end chunk of the current segment and the start chunks of possible following segments. The synthesis quality difference between random segment shuffling and similarity-controlled segment shuffling was evaluated with a listening test presented in Section 4.

The first segment of the texture is selected randomly from the collection of all available segments. Distances between the last chunk of the selected segment and start chunks of all segments available are extracted from the similarity matrix calculated in the analysis stage. Candidate segments are ranked by the similarity (minimum distance), and the next segment is randomly selected from the top N candidates. The parameter N controls the change rate of auditory scene, and in this work, it is fixed to 10. The procedure is shown in Figure 4. This segment-order-randomization approach will generate a homogeneous and timbrally slowly changing auditory scene, where transitions between segments are indistinguishable.

The amount of audio data per location will most likely be small due to the cost of data collection. Thus, the system has to be able to utilize a low amount of source audio efficiently and have a sensible approach for reusing the segments while avoiding noticeable repetitions. Some constraints are set for the segment selection: at each time, it is not allowed that the next segment is a repetition of the current segment (avoids stuttering) or the segment which was originally following the current segment in the source signal (forces shuffling). A time constraint is also imposed so that re-usage of a segment is allowed only after a certain waiting period. We decided on waiting period of 20 s. This controls how repetitive the resulting audio texture will be. The diversity of audio texture will depend on the length and the complexity of source signal.

The chosen segments are concatenated one after another based on the order determined by the presented segment selection rules. The segments are linearly cross-faded within 2 s, outside of their boundaries, for smooth transitions. The cross-fading process is illustrated in Figure 5.

Example of source signals and synthesized audio textures are available at www.cs.tut.fi/~heittolt/audiotexture/.

4 Subjective evaluation

Three listening tests were conducted to study the proposed method for location-specific audio textures. The first listening test studied the benefits of using similarity-controlled segment shuffling over random segment shuffling (later denoted by experiment 1). The second test studied the relative difference in quality between a synthesized audio texture and an original audio sample (later denoted by experiment 2). First and second listening tests were conducted in a forced choice paired comparison setup. The third listening test studied the absolute quality of the synthesized audio textures and their ability for representing the auditory scene of various locations (later

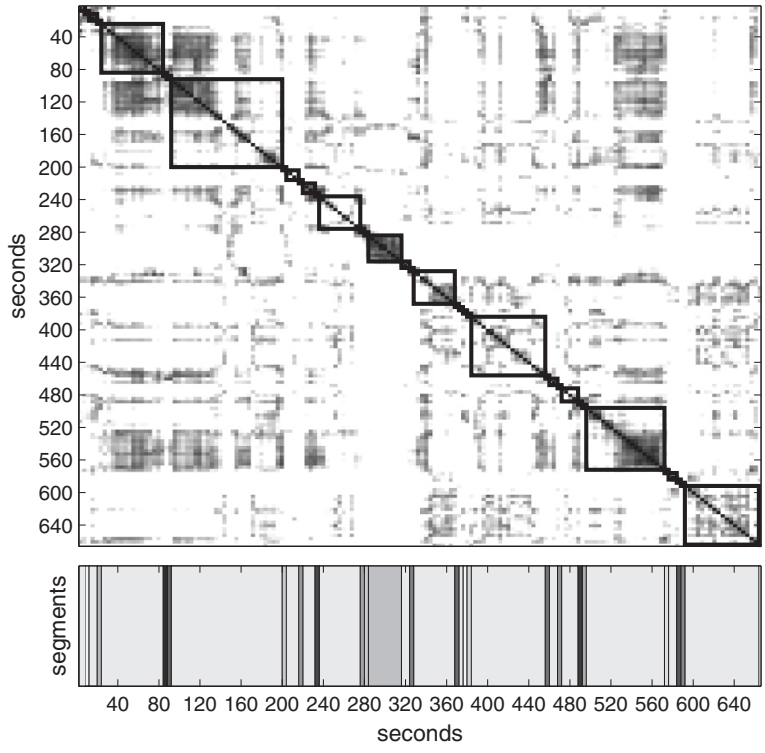

Figure 3 Example of a distance matrix (upper panel) and corresponding segmentation (lower panel). In the lower panel, the segments belonging to the same cluster are marked with the same color. The corresponding segmentation is marked on the distance matrix by the rectangles.

denoted by experiment 3). This test was conducted in an absolute rating test setup.

The tests had two goals. The first goal was to study the quality of the textures: can people notice a degradation in audio quality when they are played an automatically created texture of an environmental recording? The second goal was to study the representativeness of the automatically created textures: can an automatically created texture credibly represent a location, such as a street, even if it had some noticeable audio quality problems? These experiments studied the audio textures from two perspectives: on one hand, the audio textures were compared directly to reference samples in a forced choice test, and on the other hand, the audio textures were presented without a reference asking subjects to rate the representativeness for a certain location. The later setup is more close to the target application of virtual location-exploration services, where audio textures are used to represent certain locations without direct reference to real recordings.

In order to allow flexibility, the listening tests were carried out in a web browser via the Internet. This allowed subjects to do the test effortlessly when they had time for it. There was no requirement for the used equipments; however, the subjects were encouraged to listen to the audio samples with high-quality headphones. At the start of the experiment, they were given details about the experiments and put through a familiarization stage.

The experiment consisted of task pages presented sequentially to the test subject, each page containing one pair of samples or an individual sample depending on the test setup. The subjects controlled the playback of the audio sample with the playback button on the task page. It was emphasized to the subjects that it is important to listen through the whole audio samples before answering the questions. After answering the given question, the subject could proceed to the next task page by clicking a button. The order of the task pages was random for all subjects.

4.1 Audio samples

The audio data used in the experiments consists of recordings from different indoor and outdoor locations in the city of Tampere, Finland. The recordings were made using a binaural setup, with the recording person wearing the microphones in his ears during the recording. During the recording, some movement was permitted inside small radius of the location and the recording person was instructed to talk as little as possible. The recording equipment consists of a Soundman OKM II Klassik/studio A3 electret microphone (Berlin, Germany) and a Roland Edirol R-09 wave recorder (Hamamatsu, Japan). The samples were recorded at a sampling rate of

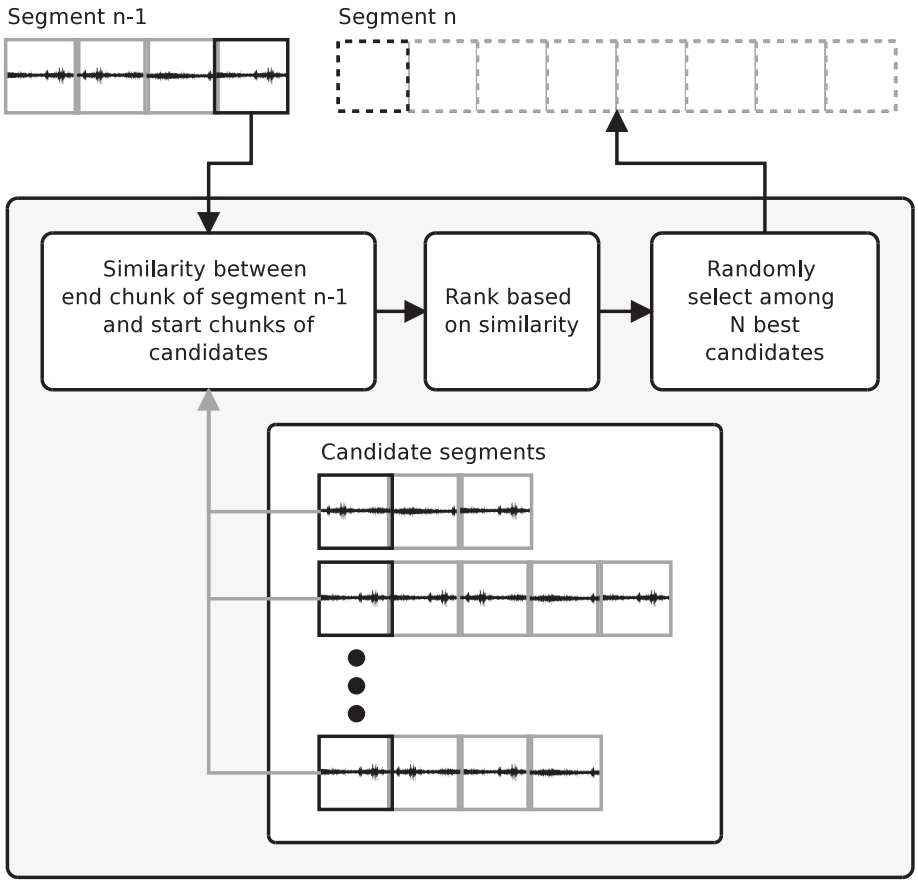

Figure 4 Similarity-controlled shuffling, finding candidate segments for the randomization.

44.1 kHz and 24-bit resolution. The stereo signals were averaged into single-channel signals for the experiments.

The amount of locations used in the listening tests was limited to four. This allowed to include sufficient amount of recordings per location to get statistically reliable results while keeping the length of the listening test per test subject reasonable. Selected locations were a pub, a restaurant, a street, and a track and field stadium during an athletics competition. These locations were selected considering that such public locations would be interesting for a virtual location-exploration service. From each of the selected locations, four audio recordings which durations ranged from 3 to 10 min were included in the listening tests, considering that this would be the typical length in crowdsourced material. The audio ambiance for the pub location contains mostly music and speech babble background, with occasional conversation in foreground. The restaurant location is a self-service type of student restaurant, located in a noisy hallway, ambiance composed mostly of background sounds like speech babble and sounds of cutlery and trays. The audio ambiance of the street location contains sound of cars passing by

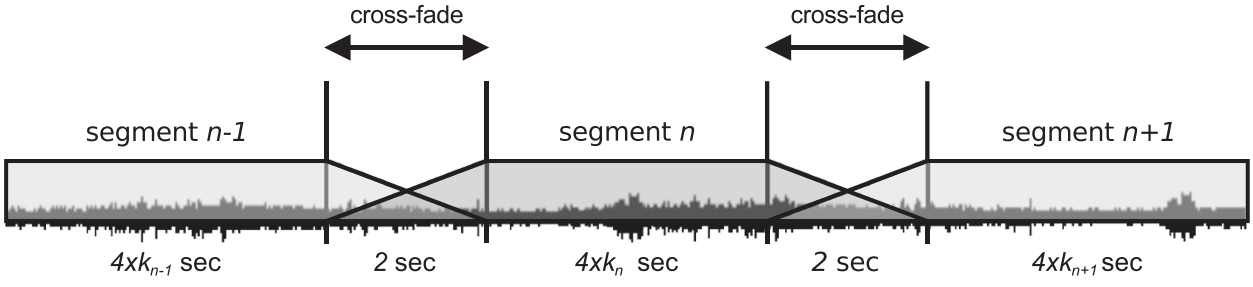

Figure 5 Synthesizing the audio from the segments; each segment comprises *k* chunks of length of 4 s.

and footstep sounds of pedestrians, with general distant traffic noise in the background. Track and field recordings were collected during the athletics competition. The audio ambiance of this location includes foreground events such as announcements and sound from the competitions on foreground and sounds of audience in the background.

4.2 Pair comparison tests

Experiments 1 and 2 used a forced choice setup, where test subjects were asked to select the more fluent and natural audio sample out of two presented samples. An example of a task page used in the tests can be seen in Figure 6. Test subjects were asked to select one sample out of a pair, based on the statement: 'Select the more fluent and natural audio sample. Audio samples are representing X location.' Before the listening test, the test subjects were guided to evaluate samples according to following cues: 'Is the order and continuity of the events in the sample natural? (fluent and natural sample), Or are there audible abrupt changes in the sound which break the fluent structure?'

The test setup for the experiments was chosen to force the test subjects to mark the preference even when there is no obvious difference between the samples. The similar setup of the two experiments allowed them to be implemented as a single test session. Pairs of samples for experiments 1 and 2 were presented in random order. This also prevents the subjects to adapt their choices during the test. The results of the two experiments were analyzed independently.

In experiment 1, a pair of samples included audio textures synthesized with two different methods, based on the same original recording. One of the samples was synthesized using the proposed similarity-controlled segment shuffling, and the other one was synthesized using a random segment shuffling approach. A 30-s excerpt was randomly selected from each audio texture. In order to ensure a fair comparison of the methods, only excerpts having at least three transitions between audio segments were selected. In experiment 2, a pair of samples included one audio texture synthesized using the proposed similarity-controlled segment shuffling and one real audio sample. A real audio sample with a length of 30 s was randomly selected from an original recording, and an audio texture was created based on the same recording. A 30-s excerpt was selected from the texture the same way as in experiment 1. The total length of test session (experiments 1 and 2) was approximately 40 min per subject, containing a total of 32 sample pairs, 16 pairs for each experiment. In total, 27 voluntary test subjects participated in the experiments 1 and 2. There were 22 male and 5 female test subjects aged between 20 and 40. People with audio research background and people without any special knowledge about audio signal processing were included in the listening tests. However, we consider these aspects of the test subjects as not having any influence on the results of the listening tests.

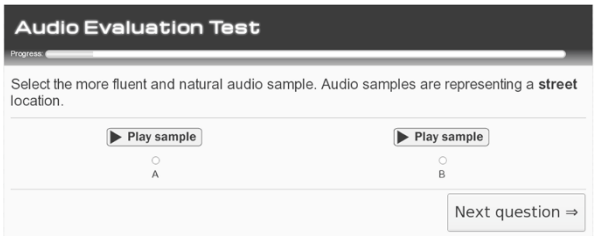

Figure 6 Caption of the experiments 1 and 2 as presented to the subjects.

4.3 Absolute rating test

The absolute quality and representativeness of the synthesized audio textures were studied in experiment 3. Real audio samples and synthesized audio textures produced by the proposed method were presented in random order to test subjects.

The listening test included four original recordings from each of the four selected locations. One audio texture was created based on each original audio recording, resulting in four audio textures available for each of the four locations. For each test subject, a 30-s excerpt was randomly extracted from each of the above audio signals. The length of the listening test was approximately 40 min per subject, containing a total of 32 excerpts (four textures and four original samples per each of the four locations). In order to ensure a fair rating, audio texture excerpts having at least three transitions between audio segments were selected, like in experiments 1 and 2.

The test consisted of several task pages, which were sequentially presented to the subject. An example of a task page can be seen in Figure 7, consisting of a single audio sample accompanied by the rating tickers for two statements:

- The audio quality of this audio sample is good.
- This audio sample sounds like it has been recorded at location X.

The ratings of the first statement provided information regarding the audio quality of the textures. The second statement simply asked how well the audio sample represents a certain type of location according to the subjects. This was used to reveal how realistic the audio textures were. The answers were given on a discrete 9-point scale, where 1 means disagree and 9 means agree.

In the familiarization stage of the listening test, an example of poor quality audio texture was presented. The example was created by random shuffle of the segments

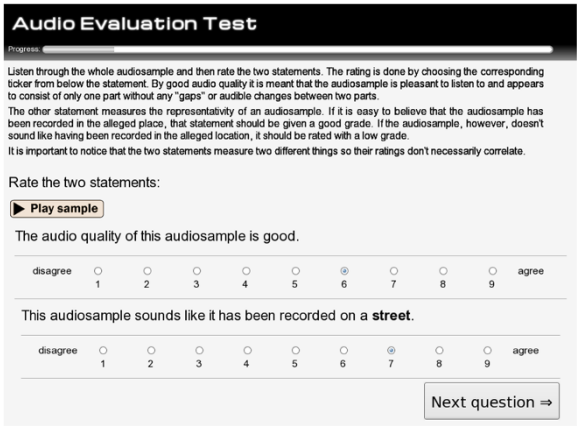

Figure 7 Caption of experiment 3 as presented to the subjects.

with no similarity-based segment selection or cross-fading while concatenating the segments. The segment transitions were easily audible in the example. The subject was informed that the example audio sample is such that it would receive the lowest ratings in terms of audio quality. A second sample was presented, this time a fragment of a real recording, and the subject was told that this example is such that it would receive the highest ratings in terms of audio quality.

The representativeness of an audio sample was explained textually: an audio sample represents a certain type of location well if it sounds as if it has been actually recorded there. It was also emphasized that an audio sample may have bad audio quality but still good representativeness for a certain location, so the two aspects of the evaluation should be kept separate.

In total, 21 voluntary test subjects participated in the listening experiment 3. Test subjects were different than in experiments 1 and 2, preventing test subjects from using prior knowledge of the signals. There were 13 male and 8 female test subjects aged between 17 and 40. As in experiments 1 and 2, both audio researchers and people without any special knowledge about audio signal processing were included in the listening test. However, we consider these aspects of the test subjects as not having any influence on the results of the listening tests.

5 Results and discussion

The results and statistical analysis of the implemented listening tests are presented in this section. Pair comparison tests, experiments 1 and 2, were implemented as one listening test session, but their results are analyzed separately.

5.1 Experiment 1

Experiment 1 studied the effect of segment shuffling method to the perceived fluency and naturalness of the synthesized audio texture. In the test, two samples were presented to test subjects: a reference sample synthesized by random segment shuffling and a sample synthesized by similarity-controlled segment shuffling. The expected outcome of this experiment was that the similarity-based segment shuffling is preferred because the transitions between segments should be concealed by similar audio content on either side of the cuts.

The results were analyzed using binomial test with a significance level of 5%, under the null hypothesis that the two options are equally likely to be selected. If the p-value is below the chosen significance level, this null hypothesis is rejected. This means that the choice is biased, so one option is more likely to be selected by the test subjects. The results are given in Table 1. There are 108 data points for each location, totaling 432 data points for the whole experiment. The 95% confidence interval and p-values resulted from the binomial test for each case are given in the table. The results indicate that the proposed method - the similarity-controlled segment shuffling-based synthesis - is able to outperform purely random segment shuffling in three out of four locations. The proposed method was preferred over the reference method in 61.1% of the overall cases, which is statistically significant.

There is significant difference in user preference for the track and field, street, and pub locations. In these locations, there are distinct foreground events - such as announcements, crowd sounds, cars passing by, conversation, which when segmented and randomly shuffled can result in very abrupt changes between adjacent segments. The restaurant location in the experiments is a self-service restaurant located in a noisy hallway with indistinguishable sounds of cutlery, trays, and speech babble forming seemingly static background ambiance, so it is understandable that timbre similarity does not offer much advantage. Based on the outcome of this test, we consider that for most locations, the similarity-controlled segment shuffling results in a more fluent and natural audio texture than a random shuffling of the segments.

Table 1 Average preference percentage and 95% confidence intervals for binomial test for experiment 1

Location	Reference preferred (%)	Proposed preferred (%)	p-value
Pub	39.8 ± 9.6	60.2 ± 9.6	0.0214
Restaurant	41.7 ± 9.6	58.3 ± 9.6	0.0507
Street	38.9 ± 9.5	61.1 ± 9.5	0.0132
Track and field	35.2 ± 9.4	64.8 ± 9.4	0.0013
Overall	38.9 ± 4.7	61.1 ± 4.7	0.0000

Reference method is random segment shuffling, and proposed method is similarity-controlled segment shuffling.

5.2 Experiment 2

Experiment 2 presented test subjects with a forced choice between an original audio sample and a synthesized one. Similar to experiment 1, the subjects were asked to choose the more fluent and natural sample. The results of the listening test are given in Table 2, together with p-values resulted from a binomial test with a significance level of 5%, under the null hypothesis that the two options are equally likely to be selected.

Overall test result for all cases shows that the null hypothesis must be rejected, and the test subjects are biased towards selecting the real recording. On smaller sets corresponding to individual locations, however, the user preference for the real recording is highlighted only for the pub location, while for restaurant, street, and track and field, the resulting p-values do not allow rejection of the null hypothesis. For the pub location, the presence of music might cause the difference in preferences as our method does not consider in any way musical structure, and while timbre might be similar between adjacent segments, music will certainly get fragmented. In other situations, it seems that the similarity-controlled shuffling is able to create sequences of segments that are realistic.

5.3 Experiment 3

In experiment 3, we presented the test subjects with an audio sample and an associated location. In this case, there was no given audio reference, therefore, no comparison to be made. The only thing the subjects had to take into account is how well the given audio characterizes the given location, from two points of view: audio quality and representativeness, scored on a scale from 1 to 9. This test refines the findings from experiment 2, by allowing subjects to rate audio quality and representativeness of a given audio for a given location based on personal experience from similar locations. The mean values of both measured aspects are listed in Table 3.

The ratings were analyzed using the statistical t-test with a 5% significance level. The null hypothesis for the t-test is that the two sets of ratings (audio sample or audio texture) have the same mean. Separate tests were performed for audio quality and representativeness. According to the results presented in Tables 4 and 5, p-values are higher than the chosen significance level. As consequence, the null hypothesis cannot be rejected.

Table 2 Average preference percentage and 95% confidence intervals for binomial test for experiment 2

Location	Reference preferred (%)	Proposed preferred (%)	p-value
Pub	63.0 ± 9.5	37.0 ± 9.5	0.0045
Restaurant	46.3 ± 9.8	53.7 ± 9.8	0.8067
Street	58.3 ± 9.6	41.7 ± 9.6	0.0507
Track and field	57.4 ± 9.7	42.6 ± 9.7	0.0743
Overall	56.2 ± 4.8	43.8 ± 4.8	0.0053

Reference method is original recording, and proposed method is similarity-controlled segment shuffling.

In this case, the result means that there is no statistically significant difference between the two sets. We cannot conclude that they are equal; however, we can perform an equivalence test to check if they are equivalent to a certain acceptable threshold. To test for equivalence, we perform two one-sided t-test (TOST) [31].

TOST has been designed specifically for bioequivalence testing of pharmaceutical products [32] and has been expanded into broader applications [33]. TOST has as null hypothesis that the difference of the two means is outside an interval. This interval is defined by the equivalence threshold θ, such that when the difference of means of the two sets is within $[-\theta, +\theta]$, the two sets are considered equivalent. The null hypotheses H_{01} and H_{02} considered in TOST are the following:

$$\begin{aligned} H_{01}: &\quad \mu_1 - \mu_2 > \quad \theta \\ H_{02}: &\quad \mu_1 - \mu_2 < \quad -\theta \end{aligned} \tag{1}$$

The choice of the equivalence threshold θ is well defined in bioequivalence testing, but for other disciplines, there are no guidelines for choosing it. The real recordings represent a control set, and in our analysis, we consider a threshold around the mean of scores given for the control set by the subjects. We performed two one-sided t-tests according to the two null hypotheses in Equation 1, for an equivalence threshold θ_{10} which is 10% of the mean of scores of the control set. The equivalence threshold is selected so that the equivalence interval at the middle of the grading scale equals to one unit.

The p-values resulted from this test, and the corresponding equivalence thresholds for each performed test are presented in Tables 4 and 5.

For a significance level of 5%, the overall set (sample size 336) passes the test both in audio quality and representativeness; as in both tests, the obtained p-value is smaller than the chosen significance level, allowing us to reject the null hypotheses. The tests performed for each separate location, however, do not allow us to reject the null hypotheses in majority of the cases. In this case, we cannot show a statistically significant equivalence at the given threshold. This can happen in experiments which have a small sample size (84 in this case), as the sample size, standard error, and equivalence threshold are related. In normal analysis situations, the equivalence threshold is chosen based on prior knowledge of the experiment and its intended application. When considering the audio quality for example, the size of the difference which can be considered to represent equivalence is not something intuitive - it is difficult to explain what an audio quality of 6.8 is and if it is equivalent to an audio quality of 7.1 or

Table 3 The mean values along with standard error of the ratings of audio quality and representativeness in experiment 3

	Audio quality		Representativeness	
Location	**Audio texture**	**Audio sample**	**Audio texture**	**Audio sample**
Pub	5.54 ± 0.24	5.93 ± 0.24	6.48 ± 0.24	6.57 ± 0.24
Restaurant	5.90 ± 0.22	5.83 ± 0.27	6.83 ± 0.21	6.86 ± 0.23
Street	5.73 ± 0.26	5.54 ± 0.25	6.87 ± 0.24	7.08 ± 0.23
Track and field	5.82 ± 0.27	6.12 ± 0.24	6.70 ± 0.27	6.57 ± 0.27
Average	5.75 ± 0.12	5.85 ± 0.13	6.72 ± 0.12	6.77 ± 0.12

not. For a complete analysis of our results, we determined at what threshold θ_c we would be able to conclude equivalence in each separate case, based on the critical t-value for a significance level of 5%. The obtained values are also presented in Tables 4 and 5.

5.4 Discussion

Based on the result of experiments 1 and 2, it seems that the similarity-controlled segment shuffling produces more fluent and natural audio texture than random shuffling; however, when given a choice, subjects selected the real recording over the audio texture. An equivalence between quality and representativeness grades for audio texture and audio samples can be concluded to certain degree.

For representativeness, the equivalence can be concluded for smaller threshold than for audio quality, as shown in experiment 3. This is understandable, since an audio texture is based on a real recording from the given location. Audio quality is of course affected by the perceived continuity of the auditory scene. In some cases, the sequentiality of events is broken by the shuffling, even if timbre continuity is considered: announcements in track and field and speech and music in pub. In such cases, for equivalence, we need to tolerate a larger threshold.

The restaurant location has overall the smallest equivalence threshold - we also noticed from experiments 1 and 2 that for this location, there was no obvious user preference for either option (random or similarity-based shuffling, method, or real recording). On the other hand, the track and field and the street locations have clear foreground sounds, and we can demonstrate equivalence only if we tolerate a one-unit threshold.

The problem of selecting an equivalence threshold beforehand for listening tests is difficult since there are no intuitive values when dealing with opinion scales. By using multiple types of experiments, we could however make some observations.

According to the results, the generated audio has comparable representativeness to real recording and a satisfactory quality. As a method for creating location-based audio ambiance, it is certainly a viable option, as it will avoid recording of large amounts of data or perceived repetitions if using smaller amount of audio. The audio texture has satisfactory properties for the proposed location, and users are likely to accept it as an audio ambiance for the location. The results showed that when facing a forced choice, in most cases, people prefer the real recording, but for the envisioned application, this is not the case. When a user is presented with audio ambiance of a location, he or she should not be bothered by artifacts or notice that it is artificially created audio, and our method was shown to fulfill these requirements.

6 Conclusions

We presented a method for audio texture creation for virtual location-exploration services. A short audio signal recorded at the given location is used to synthesize a

Table 4 The *p*-values obtained from performing location-wise *t*-tests and TOST with a 10% equivalence threshold θ_{10} for audio quality results from experiment 3

Location	*t*-test	TOST	θ_{10}	Equivalence with threshold
Pub	0.2359	0.5456	0.59	$\theta_c \leq 1.04$
Restaurant	0.8351	0.1372	0.58	$\theta_c \leq 0.75$
Street	0.5976	0.3152	0.55	$\theta_c \leq 0.90$
Track and field	0.4268	0.4014	0.61	$\theta_c \leq 1.04$
Overall	0.5425	0.0067	0.59	$\theta_c \leq 0.45$

According to the *t*-test, none of the cases has statistically significant difference. According to TOST, equivalence cannot be concluded for any location with θ_{10}.

Table 5 The *p*-values obtained from performing location-wise *t*-tests and TOST with a 10% equivalence threshold θ_{10} for representativeness results from experiment 3

Location	*t*-test	TOST	θ_{10}	Equivalence with threshold
Pub	0.7756	0.0939	0.66	$\theta_c \leq 0.75$
Restaurant	0.9385	0.0332	0.69	$\theta_c \leq 0.63$
Street	0.5161	0.1354	0.71	$\theta_c \leq 0.86$
Track and field	0.7325	0.1708	0.66	$\theta_c \leq 0.89$
Overall	0.7655	0.0002	0.68	$\theta_c \leq 0.38$

According to the *t*-test, none of the cases has statistically significant difference. According to TOST, equivalence can be concluded only for the restaurant location with θ_{10}.

new arbitrary-length signal, while conserving location-specific characteristics of the source audio. The proposed audio texture creation method consists of two main steps: analysis and synthesis. In the analysis stage, the source audio recording is segmented into homogeneous segments. In the synthesis stage, the segments are rearranged and concatenated to create the audio texture. The segment sequence is created by selecting next segment to the end of sequence randomly among timbrally most similar segments.

Three listening experiments were conducted to assess audio quality of the created audio textures. Forced choice experiments showed that the proposed method is better than random shuffling of segments, but user preference is biased toward the real recordings. With the absolute rating experiment, we showed that representativeness and audio quality of the created audio textures are comparable within certain threshold to real recordings. This means that automatically created audio textures could be used in a virtual location-exploration service in a credible way instead of original audio recordings. If the events in the source auditory scene are sequential, the resulting audio texture will most likely break the order of events and fail in being realistic. This should not be a problem for outdoor environments but can be annoying in sport events or music concerts where sound events might happen in a specific order.

There are several directions for future work. To increase the quality of the created audio textures for ambiances containing music, a music detector front-end could be used to detect sections containing music and perform the rearranging of the audio segments by following music-specific rules such as rhythm and structural continuity in addition to general timbral similarity. In addition, estimation of the auditory scene complexity could help in determining the number of clusters which should be used in the modeling of segments.

In addition, methods for creating audio textures from more than one recording should be studied. Nowadays, it is easy to obtain crowdsourced data; this would be an excellent source of audio content for such a system. A larger set of source recordings would allow creating more versatile content compared to rearranging segments from a single recording. However, this would also need paying attention to possible acoustical mismatches between recordings, e.g., recorded at different time of the day.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was financially supported by the Academy of Finland under the grant 265024 (Heittola and Virtanen). This work was partially supported from the grant 251170 (Mesaros) Finnish Centre of Excellence Program (2012-2017).

Author details
[1]Department of Signal Processing, Tampere University of Technology, P.O. Box 553, Tampere 33101, Finland. [2]Department of Signal Processing and Acoustics, Aalto University, P.O. Box 13000, Aalto 00076, Finland. [3]Nokia Research Center, Visiokatu 3, Tampere 33720, Finland.

Received: 14 December 2012 Accepted: 4 March 2014
Published: 11 March 2014

References

1. Nokia, Here Maps 3D (2013). http://here.com/3D, Accessed 15 November 2013
2. Google, Google Maps with Street View (2013). http://www.google.com/streetview, Accessed 15 November 2013
3. Microsoft, Microsoft Streetside (2013). http://www.microsoft.com/maps/streetside.aspx, Accessed 15 November 2013
4. J Vroomen, Bd Gelder, Sound enhances visual perception: cross-modal effects of auditory organization on vision. J. Exp. Psychol.: Human Perception and Performance. **26**(5), 1583–1590 (2000)
5. F Frassinetti, N Bolognini, E Ladavas, Enhancement of visual perception by crossmodal visuo-auditory interaction. Exp. Brain Res. **147**(3), 332–343 (2002)
6. JB Krygier, Visualization in modern cartography, in *Sound and Geographic Visualization* (Pergamon Press, London, UK, 1994), pp. 149–166
7. R MacVeigh, RD Jacobson, Increasing the dimensionality of a geographic information system (GIS) using auditory display, in *Proceedings of the 13th International Conference on Auditory Display* (McGill University, Montreal, Canada, 2007), pp. 530–535
8. J Schiewe, AL Kornfeld, Framework and potential implementations of urban sound cartography, in *12th AGILE International Conference on Geographic Information Science* (AGILE, Hannover, Germany, 2009)
9. A Eronen, V Peltonen, J Tuomi, A Klapuri, S Fagerlund, T Sorsa, G Lorho, J Huopaniemi, Audio-based context recognition. IEEE Trans. Audio, Speech, and Language Process. **14**, 321–329 (2006)
10. S Chu, S Narayanan, CC Kuo, Environmental sound recognition with time-frequency audio features. IEEE Trans. Audio, Speech, and Language Process. **17**(6), 1142–1158 (2009)
11. D Korpi, T Heittola, T Partala, A Eronen, A Mesaros, T Virtanen, On the human ability to discriminate audio ambiances from similar locations of an urban environment. Personal and Ubiquitous Comput. **17**(4), 761–769 (2013)
12. L Lu, L Wenyin, HJ Zhang, Audio textures: theory and applications. IEEE Trans. Speech and Audio Process. **12**(2), 156–167 (2004)
13. D Menzies, Physically motivated environmental sound synthesis for virtual worlds. EURASIP J. Audio, Speech, and Music Process. **2010**, 1–11 (2010)
14. S Kersten, P Purwins, Sound texture synthesis with hidden Markov tree models in the wavelet domain, in *Sound and Music Computing Conference* (SMC, Barcelona, Spain, 2010)
15. D Schwarz, ed. by G Peeters, State of the art in sound texture synthesis, in *Proceedings of the 14th International Conference on Digital Audio Effects (DaFx-11)* (IRCAM–Centre Pompidou, IRCAM, Paris, France, 2011), pp. 221–231
16. N Finney, J Janer, Soundscape generation for virtual environments using community-provided audio databases, in *W3C Workshop: Augmented Reality on the Web* (Barcelona, Spain, 2010)
17. T Jehan, Creating music by listening. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2005
18. A Zils, F Pachet, Musical mosaicing, in *Proceedings of the COST G-6 Conference on Digital Audio Effects (DaFx-01), Volume 2* (University of Limerick, Limerick, Ireland, 2001), pp. 39–44
19. McJ Dermott, A Oxenham, E Simoncelli, Sound texture synthesis via filter statistics, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (IEEE Signal Processing Society, New Paltz, NY, USA, 2009), pp. 297–300
20. D Schwarz, Corpus-based concatenative synthesis. IEEE Signal Process. Mag. **24**(2), 92–104 (2007)
21. AJ Hunt, AW Black, Unit selection in a concatenative speech synthesis system using a large speech database, in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE Signal Processing Society, Atlanta, GA, 1996), pp. 373–376

22. Sound Around You (2013). http://www.soundaroundyou.com, Accessed 15 November 2013
23. C Mydlarz, I Drumm, T Cox, Application of novel techniques for the investigation of human relationships with soundscapes, in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings, Volume 2011* (Institute of Noise Control Engineering, Osaka, Japan, 2011), pp. 738–744
24. J Freeman, DiC Salvo, M Nitsche, S Garrett, Soundscape composition and field recording as a platform for collaborative creativity. Organised Sound. **16**(3), 272–281 (2011)
25. S Innami, H Kasai, Super-realistic environmental sound synthesizer for location-based sound search system. IEEE Trans. Consumer Electron. **57**(4), 1891–1898 (2011)
26. T Heittola, A Mesaros, A Eronen, T Virtanen, Context-dependent sound event detection. EURASIP J. Audio, Speech, and Music Process. **2013**(1-13), 1 (2013)
27. M Figueiredo, J Leitão, A Jain, ed. by E Hancock, M Pelillo, Berlin on fitting mixture models, in *Energy Minimization Methods in Computer Vision and Pattern Recognition, Volume 1654 of Lecture Notes in Computer Science* (Springer Berlin Heidelberg, Germany, 1999), pp. 54–69
28. T Virtanen, M Helen, Probabilistic model based similarity measures for audio query-by-example, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (IEEE Signal Processing Society, New Paltz, NY, USA, 2007), pp. 82–85
29. J Foote, Automatic audio segmentation using a measure of audio novelty, in *IEEE International Conference on Multimedia and Expo, Volume 1* (IEEE Computer Society, New York, NY, USA, 2000)
30. A Jacobs, ed. by G Antoniou, G Potamias, C Spyropoulos, and D Plexousakis, Using self-similarity matrices for structure mining on news video, in *Advances in Artificial Intelligence, Volume 3955 of Lecture Notes in Computer Science* (Springer Berlin Heidelberg, Berlin, Germany, 2006), pp. 87–94
31. DJ Schuirmann, A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J. Pharmacokinetics and Biopharmaceutics. **15**(6), 657–680 (1987)
32. GB Limentani, MC Ringo, F Ye, ML Bergquist, EO McSorley, Beyond the *t*-test: statistical equivalence testing. Analytical Chem. **77**(11), 221–226 (2005)
33. BL Stegner, AG Bostrom, TK Greenfield, Equivalence testing for use in psychosocial and services research: an introduction with examples. Evaluation and Program Plann. **19**(3), 193–198 (1996)

doi:10.1186/1687-4722-2014-9
Cite this article as: Heittola *et al.*: **Method for creating location-specific audio textures.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:9.