Designing Sound Collaboratively - Perceptually Motivated Audio Synthesis

Niklas Klügel
Technische Universität München
Department of Informatics
Boltzmannstrasse 3
85748 Garching bei München
kluegel@in.tum.de

Timo Becker
Ludwig-Maximilians-Universität München
Department of Informatics
Amalienstrasse 17
80333 München
timo.becker@campus.lmu.de

Georg Groh
Technische Universität München
Department of Informatics
Boltzmannstrasse 3
85748 Garching bei München
grohg@in.tum.de

ABSTRACT

In this contribution[1], we will discuss a prototype that allows a group of users to design sound collaboratively in real time using a multi-touch tabletop. We make use of a machine learning method to generate a mapping from perceptual audio features to synthesis parameters. This mapping is then used for visualization and interaction. Finally, we discuss the results of a comparative evaluation study.

Keywords

Collaborative Music Making, Sound Design

1. INTRODUCTION

Sound design is key to modern electronic music composition and performance as it is a form-bearing dimension of music [11]. Endogeneous motivators for this creative task encompass the goal oriented creation of sound forms with functional intent (e.g. communicating the compositional structure) or the explorative creation of sound forms..

Technically, various methods to generate various timbres exist, such as Abstract-, Physical- and Spectral Synthesis or via processing recordings. In general, the parameters for the synthesis method are not grounded in the *perceptual* but rather in the *structural* domain. Hence, they may frequently exhibit non-linear behavior and inter-dependencies regarding the perceptual qualities of the output. Changing timbre is therefore non-intuitive without prior knowledge about this structure and its technical functioning [23, 24], especially for novices. In a collaborative setting, this may be even more problematic as tasks involving group creativity greatly benefit from heterogeneous groups including novices [9, 19, 28] to foster social creativity. This beneficial effect is further amplified by the phenomenon of *Flow* [5] which stimulates the implicit learning process, enables empathic involvement with the music [21] and is also bound to the perception of *social presence*.

Group Flow as a social experience has been shown to foster more valuable musical results [17] and is a key success factor in Computer Supported Collaborative Music-Making (CSCM) [27] and thus acts as an intrinsic motivator for the group. In that role, it also has the effect of supporting the learning process for musical expression, thus contributing to mediating the interaction with the shared CSCM environment and, finally, the social interaction with peer members. We thus regard the integration of these social effects as highly beneficial for our task. The utilization of a multi-touch table especially offers means to facilitate social communication protocols [29, 12] and we will therefore use such a device in this contribution.

Certain aspects of Human Computer Interaction (HCI) have been identified as important for collaborative applications. Based on the work by Cockburn [3] and Dourish [6], the most prevalent ones for our task are: *Group Awareness*, (ability of individuals in the group to share and gain mutual knowledge about the ongoing activities) and *Group Articulation*, (ability to partition the group task into units that can be combined and substituted). Fostering these aspects will serve as basic requirement for our interaction design.

Within this contribution, we use the term *timbre* as an abstract descriptor of perceptually salient acoustic features (e.g. low level such as frequency components or high level such as the spectral centroid) that can be modeled and measured by IT systems.

2. RELATED WORK

W The majority of research related to mapping from timbre to synthesis parameters focuses on re- or cross-synthesizing the original sound sources from timbral features. For our use-case, however, we are interested in applying this mapping foremost as a paradigm for interaction, so the main issue is human cognitive manageability. This field of HCI has not been studied extensively [26].

The most relevant environment for this contribution that utilizes cross-synthesis is CataRT [25], not because of the process technically synthesizing timbre (concatenative synthesis) but since the core concept for interaction is the navigation in a timbre space. For this, CataRT and subsequent developments use the high-level timbral features as orthogonal axes directly. This means that only a low (2-3) number of features span this space, discarding timbral qualities that may add valuable information and possibly reducing the representative quality of the space since features may be correlated.

To circumvent these problems, CataRT allows the user to re-define the space at run-time. For the collaborative use-case, this introduces the conceptual challenge how a shared navigational space can have user-specific views without hampering awareness and articulation. Nuvolet [4], which is based on CataRT, uses a static space for this case. It aims to support a collaborative, virtuosic performance enhancing the performer-to-audience relationship. It thus lacks features for the collaboration itself with respect to shareability and awareness. The cognitive load and imprecision initially induced by the camera-based 3D control method are also problematic for novices.

Apart from high-level timbral features, also low-level, di-

[1]We submitted a longer, more comprehensive version of this paper to `arxiv.org`

NIME'14, June 30 – July 03, 2014, Goldsmiths, University of London, UK.

mensionality reduced features can be used to span the timbre space as in [20, 16]. However, the resulting timbre space may not have an obvious relationship to the human perception [16], thus further processing steps are necessary. Moreover, any mapping to synthesis parameters *and* the interrelationship between timbral features are bound to be highly non-linear [10]. Non-linear mappings such as Self-Organizing Maps (SOMs) are therefore preferable and have been already applied to sound design. Especially the approach in [7] overlaps with our use-case as the generated 2D-SOM representation of the timbre-space is used as interaction metaphor. Due to beneficial mathematical properties, we will use Generative Topographic Mapping (GTM) instead of SOM (see section 5). Furthermore, we'll use high-level features since otherwise we would have the additional problem of defining a proper mapping that is computationally expressive enough to abstract from low-level features to perceptually more meaningful ones. To conclude, we see the collaborative use-case of designing sound a largely untapped territory for fostering creative endeavors.

3. SYNTHESIS & CORPUS

We need to be able to synthesize a large variety of different timbres in order to not severely limit expressivity. It is the goal to generate a corpus $\mathcal{S}$ such that each sound $s_i \in \mathcal{S}$ can be analyzed for its timbral features $t_i \in \mathcal{T}$ yielding the mapping $s_i \leftrightarrow t_i$. Later on, we would like to synthesize s_i again in real time within a tonal context.

We opted to develop our own synthesis model such that all parameters of the synthesis (e.g. pitch) are known instead of using re-synthesizing recorded sounds. Since it is non-trivial to analyze all time-frequency relationships reliably for an *arbitrary* sound source, re-synthesizing a sound may render its original timbre-space relationship invalid. To create the corpus it is therefore necessary to sample various parameter settings $p_i \in \mathcal{P}$ from the possible parameter combinations $\mathcal{P}$ to generate the sounds s_i for the analysis. Hence, we are interested in having a low number of parameters. We used Vector Phase Shaping (VPS) [13] at the core of our synthesis model since it allows to create various waveforms and filter phenomena with only two parameters per oscillator. Furthermore, our synthesizer model uses two oscillators that can be switched to various master-slave configurations. The sounds generated can evolve over time by modulating involved synthesis parameters using triggered envelopes. To blur or emphasize spectral peaks we added a flexible effects chain (chorus etc.). With this configuration the synthesizer is able to create a variety of timbres such as bassy, percussive, leading and atmospheric ones.

Further compacting the parameterization led to 16 variants that were discretized to at most 20 steps, resulting in $|\mathcal{P}| \approx 10^{15}$ possible parameter combinations.

Given the cardinality of $\mathcal{P}$, it becomes clear that synthesizing all parameter configurations is not feasible to cover the whole extent of timbral varieties that can be achieved with the proposed synthesizer for the corpus. We thus used a combination of expert input and a high dimensional search method using audio similarity as heuristic. $\mathcal{P}$ forms a hypercube if we interpret the normalized parameters as orthogonal axes. A hypercube can be further split along any axis, creating two siblings with a smaller volume (less included parameter configurations). Furthermore, we can determine an estimate of the similarity of the sounds within a hypercube via the similarity for the *synthesized* audio from two parameter configurations at opposing vertices on each side of such a split. Using this heuristic, we essentially perform a multi-dimensional binary search, creating dissimilar sounds along a path that eventually leads to a recursively generated volume containing mostly similar ones. We can gain a speed-up as the produced siblings are independent and can be processed in parallel. Given a set of parameter configurations - or presets - created by an expert, the method can search for sounds in between these according to the similarity measure; each pair of presets is then interpreted as two extreme vertices describing a hypercube uniquely. We used the similarity measure proposed in [22], as its robustness to arbitrary sounds has been shown. The nature of VPS to predictably generate, for our ears, mostly musical spectral effects makes this method applicable. The length of the generated samples for the subsequent feature analysis has been set to 4 seconds.

4. FEATURES

The analysis of the corpus regarding high-level timbral features follows [14], using the MIRToolbox [15]. These generated features are largely time series data, which, in our use-case, are difficult to integrate: First technically difficult, as the synthesis uses modulators whose state would have to be saved with the generated audio in order to properly represent every frame of a time-series feature in a $t_i \leftrightarrow p_i$ mapping. Second, conceptually difficult, as path operations are now the canonical way of exploring and designing sound, thus an appropriate interface may be complex and may overburden users. And third, practically difficult, as the memory requirements for the visualization method is $\mathcal{O}(n^2)$ and our feature data set exceeds 100GB. Thus we collapsed the originally 30 dimensional time series data onto a single 368 dimensional feature vector by extracting statistical properties of the features. Since our projection method introduced below, is not well suited for this high dimensionality, we applied a greedy forward feature selection yielding 50 features. Furthermore, the original synthesis parameters were added to the feature vector in order to include to some degree information about the temporal evolution of a sound. Initial listening tests revealed a significantly positive impact on the quality of the mapping. We allot this finding to a reduction of the amount of ambiguous information for the entire feature vector, as the conversion reduced the temporal descriptiveness for some of the original features (e.g. the statistical properties of the spectral centroid for a sound being played back in forward or reverse are similar).

5. MAPPING & VISUALIZATION

A reduction of the high dimensional timbre space to two dimensions is deemed especially helpful in the context of 2D multi-touch applications. For this task we preferred Generative Topographic Mapping (GTM) [1] to SOM because of guaranteed convergence and topology preservation.

GTM defines a probability density modeling the quality of the mapping y between low L-dimensional latent variables, x_k, and high D-dimensional data-points t_i. In our case, L is the 2D visualization space and D is the feature space. The mapping y defines a non-linear transformation via weighting Gaussian basis functions. The centers of the these basis functions form a uniform grid in the latent space. So far, the intrinsic dimensionality of the mapping in data space is L. Only using this strictly confined L-dimensional manifold does not allow for some variance between the observed variables (the feature data) and the images of the latent variables. Therefore, the manifold is convolved with an isotropic Gaussian noise distribution with an inverse variance β giving it some volume. These probabilistic properties allow to evaluate the quality of the mapping, such that the parameters β and the weightings can be determined by Expectation Maximization. The continuous and smooth nature of y allows for the topology preserving nature of the GTM - neighbor points in the latent space remain neighbor points in data space. This smoothness can be controlled by the parameters of the Gaussian basis functions. At the end of the algorithm, the responsibility (probability) of each latent variable for each data point can be evaluated. As latent variables are arranged on a grid, the position of each data point projected

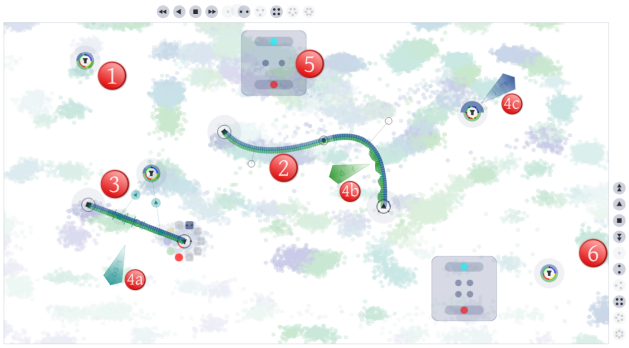

Figure 1: Screenshot of our prototype showing Nodes (1), Paths (2-3) and tools (4a-4c) as well as global controls for tonal parameters (5) and playback (6)

down can be determined by weighting the latent variables' grid positions with their responsibility. In our case it is further necessary to perform a Z-Score transformation of the feature data since the variance of the features diverge vastly (by the factor $\approx 10^9$) but the global parameter β applies to all dimensions.

To help users differentiate more easily between the clusters of points in the projection, we performed a rough color coding to indicate cluster membership using K-Means Clustering in the feature space ($K \approx 50$).

Since we know which data point is projected onto which 2D position, we can construct an element-wise bijective mapping. For interaction, in order to find the nearest projected point in latent space from an arbitrary point (e.g. a user's touch) in the 2D space in real time, K-Nearest Neighbor search with $K = 1$ using a KD-tree was applied. The resulting time-complexity of $\mathcal{O}(k \log n)$ with $\approx 10^6$ points permits comparatively quick look ups. In section 8 we will give an example of how we can benefit from more neighbors. Finally, we applied a generic hash map to associate a feature vector with the respective set of generating parameters. We will call the visualization, $p_i \rightarrow s_i \rightarrow t_i \rightarrow x_i$, and the mapping, $x_i \rightarrow t_i \rightarrow s_i \rightarrow p_i$, *Timbre Surface* for future reference.

6. PROTOTYPE

For the prototype the focus is on including novices but also experts in the collaborative sound design process (**expertise of users**), supporting both purposeful as well as explorative sound construction (**user objective**). Aiming at purposeful collaboration we deem some explanatory exposition with respect to the interaction language and expressive possibilities permissible (**situational context**). Regarding the **mode of collaboration**, we opted for the distributed leadership [2] since this allows us to evaluate whether the proposed synthesis and interaction method aid users in comparison to a more conservative approach. Besides using methodologies for fostering awareness, shareability, articulation such as different levels of task coupling (**facilitation of collaboration**), we support public and private (acoustic) spaces as they have been shown to give users more creative freedom to formulate contributions [8].

As stated earlier, users can design their sounds in a tonal context. For this, streams of harmonically fitting note events are generated preserving the validity of the timbral mapping [18]. Private acoustic spaces were supported by separate headphone output channels that can be routed freely to hear any sound being played.

The Timbre Surface is at the core of the operational design spanning the background of the UI. Furthermore, we use the concept of a visual draggable*Node* (fig. 1, item 1) to facilitate the selection and playback of sounds according to its 2D position. Nodes can be created and removed by simple gestures and manipulated to change the sound's pitch and volume.

Nodes can be connected to form Paths (fig. 1, item 2) to create sounds that change over time. This expresses a timed motion over the Timbre Surface whose slope in between two Nodes can be altered with control points of a Bezier curve. The segments of a Path are used for visualization, showing the progress of playback but also for representation of the properties duration, volume and pitch at a certain point in time. By chaining paths (fig. 1, item 3) more complex sequences of sounds can be created.

Tools can be instantiated which can be dragged and "docked" to an item to be modified (again via dragging). Figure 1, items 4a, 4b, 4c show the tools for changing volume, pitch, and duration of a segment.

A global menu can be used to change the playback of all Nodes and Paths but also to show the widgets that manipulate the note streams allowing to modify e.g. note length and frequency of an arpeggiated sequence of pre-programmed chord progressions.

7. EVALUATION

We created a data-set of 60,000 sounds, which is based on 1,500 expert presets, generated by the method described above. We conducted a one week user study, comparing our approach to a classical one and evaluating the design of the application with respect to user experience and collaboration.

7.1 Organization of the study

The study was divided into three consecutive parts: comparison, experimentation and questionnaire. 22 people took part it in groups of 2-4 collaborators (mean age 25.5, 27% female). 59% of them had used music software before, 55% played an instrument, and only 27% had used a synthesizer before. 14% had composed a musical piece and 5% designed sound. Regarding the technical knowledge, 59% had previously used a multi-touch table. The vast majority of users did not indicate any expectations towards the application or (collaborative) sound design.

For the comparative part, we developed a simple alternative application that models a classic approach to sound design in a collaborative setting, using the metaphor of a shared instrument where users can change only technical parameters of our synthesis method using simple sliders. These are grouped by functionality as widgets that can be moved and rotated freely.

After a short explanation, users were given 5 minutes to experiment with the classic approach and after that, again after an introduction, 5 minutes to experiment with our prototype. We then presented all of the remaining functionality and gave participants 25 minutes to delve deeper into the application but also to get accustomed to each other. A computer-mediated questionnaire addressing the comparison-, usability-, interaction- & information design, visual design and finally the collaboration itself. Most questions used a 5-level Likert scale.

7.2 Results

The questionnaire showed mixed results regarding the overall ease of use, mental effort and the perceived ability to *purposefully* execute ideas for our new application in comparison to the classical one. Corroborating the latter, participants were not always able to realise their musical and timbral ideas because they felt that finding specific sounds quickly was rather hard. With respect to collaboration, it was stated that sharing and continuing work of others was not sufficiently *facilitated* while it was *perceived as easy*. This means that articulation was perceived to be fostered, as opposed to shareability. Apart from these mixed results, the feedback for the application was very positive. The comparison between the classic and our approach showed that ours

was deemed more musically inspiring and incited people more to experiment and collaborate. It was further stated that it helped to obtain more musical results both alone and in the group. The relationship between input and auditive output was found to be more understandable. People perceived the new application as being able to provide more freedom in creating timbres and, regarding utility, most participants were able to find interesting timbres and create interesting complex ones. Most interestingly, a minority stated that they had been shown new vistas with respect to music and timbre. In relation to engagement and pleasure, the majority experienced Flow phenomena. Collaboratively designing and experimenting with sound were rated as providing fun while it was felt that the collaboration in general had been fostered. In terms of awareness and shareability, the possibility to experiment in a private auditory space was received very positively. Finally, the application was favorably reviewed concerning the interaction and information design, as well as aesthetics.

8. DISCUSSION & CONCLUSION

The evaluation revealed some shortcomings with respect to Navigation and Awareness. Regarding the first shortcoming, users stated that they found Timbre Surface incoherent as small changes in position did not translate to small changes in timbre and that the clusters in the projection did not always have a comprehensible inter-relationship. The quality of the Timbre Surface depends crucially on the quality of the features. In this way the issue can be remedied with a different set of timbral features or different encodings thereof. Additionally, the negative evaluation results with respect to the precision of the Timbre Surface led to a subsequent experiment investigating the influence of more thorough feature selection methods via observing the likelihoods involved in the GTM. The results showed that these methods can improve the quality of the projection significantly. The GTM method itself also provides parameters that can be further adjusted. Real-time interpolation of the sample points could lead to a smaller data-set and therefore disentangle the visualization but can also lead to a homogenization of parameter settings depending on their neighborhood. For a quick evaluation of this method we used the nearest 8 neighbors of a position and weighted their parameters according to the Euclidean distance from that position. This created a sufficiently smooth mapping where gradual changes in the 2D position led to gradual changes in the aural output. However, this removed many of the original timbres from the Timbre Surface since this linear interpolation does not inherit the non-linear nature of the GTM. Hence, a more complex interpolation method is needed which takes the GTM into account. For example, the gradient of the generated responsibilities could be used for weighting. Another approach is to introduce complementary UI tools, that can pull clusters apart (e.g. a "magnifing glass").

Our application was received as a valuable alternative to the classic approach to sound-design, being perceived as more musical, expressive, inspiring, comprehensible and inciting towards collaboration. Participants stated that they were able to find interesting and design complex timbres. This activity was perceived as pleasurable and Flow-inducing. Furthermore, the collaborative use-case is seen as fun. The application as been reviewed as supporting and fostering collaboration. Although the named shortcomings conflict with the goals set at the beginning of this paper, we do not assess them as overly severe since they are not conceptual issues but rather technical ones that can be approached systematically within further research. To conclude, given this positive feedback and the amount of committed suggestions for improvements by participants, we see the results as satisfactory. In spirit of our previous work, the data-sets (feature- & parameter data, evaluation) and source-code of our application and frameworks are available from our website.

9. REFERENCES

[1] C. Bishop et al. GTM: A principled alternative to the self-organizing map. In *Proc. Int Conf. Artif. Neur. Net.*, pages 165–170. Springer, 1996.

[2] T. Blaine et al. Collaborative Musical Experiences for Novices. *J. New Music Res.*, 32(4), Dec. 2003.

[3] A. Cockburn et al. Four principles of groupware design. *Interact. Comput.*, 7(2), 1995.

[4] J. M. Comajuncosas et al. Nuvolet : 3D Gesture-driven Collaborative Audio Mosaicing. *Proc. Int. Conf. NIME*, 2011.

[5] M. Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. Harper and Row, 1990.

[6] P. Dourish et al. Awareness and coordination in shared workspaces. In *Proc. Int. Conf. CSCW*, 1992.

[7] A. Eigenfeldt et al. Realtime timbral organisation: Selecting samples based upon similarity. *Organised Sound*, 15(2), 2010.

[8] R. Fencott et al. Computer musicking: Hci, cscw and collaborative digital musical interaction. In *Music and Human-Computer Interaction*. Springer, 2013.

[9] G. Fischer. Distributed intelligence: extending the power of the unaided, individual human mind. In *Proc. Int. Conf. Adv. Vis. Interf.* ACM, 2006.

[10] M. Hoffman et al. Feature-based synthesis: Mapping acoustic and perceptual features onto synthesis parameters. In *Proc. Int. Conf. ICMC*. Citeseer, 2006.

[11] P. Holmes. An exploration of musical communication through expressive use of timbre: The performer's perspective. *Psychol. Music*, Mar. 2011.

[12] E. Hornecker et al. From entry to access: how shareability comes about. In *Proc. Conf. Des. Pleasurable Prod. Interf.* ACM, 2007.

[13] J. Kleimola et al. Vector Phase Shaping Synthesis. *Proc. Int. Conf. DAFX*, 2011.

[14] N. Klügel et al. Towards Mapping Timbre to Emotional Affect. *Proc. Int. Conf. NIME*, 2013.

[15] O. Lartillot et al. A matlab toolbox for music information retrieval. In *Data analysis, machine learning and applications*. Springer, 2008.

[16] S. Le Groux et al. Perceptsynth: mapping perceptual musical features to sound synthesis parameters. *Proc. Int. Conf. Acoust. Speech Sig. Process.*, 2008.

[17] R. MacDonald. Creativity and flow in musical composition: an empirical investigation. *Psychol. Music*, 34(3), July 2006.

[18] J. Marozeau et. al. The dependency of timbre on fundamental frequency. *J. of Acoust. Soc. America*, 114(5), 2003.

[19] G. Morgan. Paradigms, Metaphors, and Puzzle Solving in Organization Theory. *Adm. Sci. Q.*, 25(4), 1980.

[20] C. Nicol et al. Designing sound: Towards a system for designing audio interfaces using timbre spaces. In *ICAD*, 2004.

[21] L. Nijs. *The musical instrument of natural extension of the musician*. Witwatersrand Univ. Press, 2009.

[22] E. Pampalk. A matlab toolbox to compute music similarity from audio. In *Proc. Int. Conf. ISMIR*, 2004.

[23] J.-C. Risset. The perception of musical sound. 2003.

[24] A. Röbel. Between physics and perception: Signal models for high level audio processing. *Proc. Int. Conf. DAFX*, 2010.

[25] D. Schwarz. The sound space as musical instrument: Playing corpus-based concatenative synthesis. *Proc. Int. Conf. NIME*, 2012.

[26] A. Seago. A new interaction strategy for musical timbre design. In *Music and human-computer interaction*. Springer, 2013.

[27] B. Swift et al. Engagement Networks in Social Music-making. *Proc. Int. Comp. Hum. Interact.*, 2004.

[28] B. Uzzi et al. Collaboration and Creativity: The Small World Problem. *Am. J. Sociol.*, 111(2), Sept. 2005.

[29] D. Wigdor et al. *Brave NUI World*. Morgan Kaufmann, 2011.