A structured design and evaluation model with application to rhythmic interaction displays

Cumhur Erkut, Antti Jylhä
Aalto University
School of Electrical Engineering
Dept. Signal Processing and Acoustics
PO Box 13000 Aalto FI-00076, Finland
cumhur.erkut@aalto.fi, antti.jylha@aalto.fi

Reha Dişçioğlu
Aalto University
School of Art and Design
Media Lab
PO Box 31000 Aalto FI-00076, Finland
reha.discioglu@aalto.fi

ABSTRACT

We present a generic, structured model for design and evaluation of musical interfaces. This model is development oriented, and it is based on the fundamental function of the musical interfaces, i.e., to coordinate the human action and perception for musical expression, subject to human capabilities and skills. To illustrate the particulars of this model and present it in operation, we consider the previous design and evaluation phase of iPalmas, our testbed for exploring rhythmic interaction. Our findings inform the current design phase of iPalmas visual and auditory displays, where we build on what has resonated with the test users, and explore further possibilities based on the evaluation results.

Keywords

Rhythmic interaction, multimodal displays, sonification, UML

1. INTRODUCTION

Structured approaches in design and evaluation of novel musical interfaces are rare. Even rarer are the cases that build on the evaluation of the previous design phase, and implement the insights gained from user observations in the next phase. There is a clear need for such cases if deployment is desired, to understand how the intentions of designers are perceived and utilized by the users.

Currently, the purpose and function of evaluation of musical interfaces are in focus within the NIME community [6]. While our knowledge on musical perception, cognition, and interaction is rapidly advancing, there is a lack of practice of describing which capabilities are addressed in design, how various aspects are constraining the utilization of these capabilities, and how the mappings between the human capabilities and computational modalities are aligned. Similar observations were reported in [5, 4] regarding multimodal interfaces, and a structured approach has been proposed.

In this paper, we are primarily interested in repurposing this model for NIME. We first explain this structured approach and the corresponding design and evaluation models in Sec. 2. We then frame the previous design and evaluation phase of *iPalmas*, our testbed for designing and evaluating rhythmic interaction [2], within this model in Sec. 3. We build on all of these to present our ideas for the next design phase of iPalmas in Sec. 4. We finally derive our conclusions and indicate our future work in Sec. 5.

NIME'11, 30 May–1 June 2011, Oslo, Norway.

2. DESIGN AND EVALUATION MODEL

The basic idea of our approach, illustrated in Fig. 1, is that multimodal interactive systems are designed to coordinate the human action and perception for a particular effect, subject to human capabilities and skills. Several constraints may break the design intentions in deployment. The model structurally decomposes the computer modalities, human capabilities, and evaluation issues in a way similar to how *Unified Modeling Language* (UML) structures a modeling domain. UML is a generic computational modeling approach in software development, in which the focus and primary artifacts of development are the *models* instead of *programs* [3]. The main goals are to understand the domain, express the solution in various abstraction levels in the form of *structural* and *behavioral* diagrams, and evaluate in the realm of models and prototypes.

2.1 Design model

Fig. 1 illustrates the multimodal interaction model based on UML profiling and extensions. The model is a synthesis of input and output modalities, and their integration, expressed however in the UML framework. Profiling means that the model is specialized for a particular domain, and extension means that it includes special modeling elements.

The model is based on the practical definition of multimodal systems for musical interaction, consisting of an interface and supporting application that aim to *produce* a particular *effect* on a user, with parameters shared by this effect and a (computational) modality. The effect can be sensory, perceptual, motor, or cognitive, often forming a hierarchy by causality: the perceptual effects are usually based on sensory effects, etc, all the way up to human musical capabilities.

The musical interface can employ a *simple modality*, for instance visual or auditory, or multiple modalities by integrating simple modalities, such as audio-visual, or audio-tactile. In this case, we talk about *complex modalities*. The *multimodal integration* can be done sequentially or concurrently, always within a *time-frame*. From the computer point of view, we acquire *input modalities* with sensors, e.g., microphones, or accelerometers, or input devices. Some input devices are *event-based* (a key-press or a mouse-click), while most sensors provide continuous data streams by sampling. *Streaming-based modalities* are always indicated by their sampling frequency attribute. These are specialized as *recognition-based modalities*, which are specified by a recognition *error-rate* attribute.

In some cases, a recognition-based modality can convert a streaming-based modality into an event-based modality.

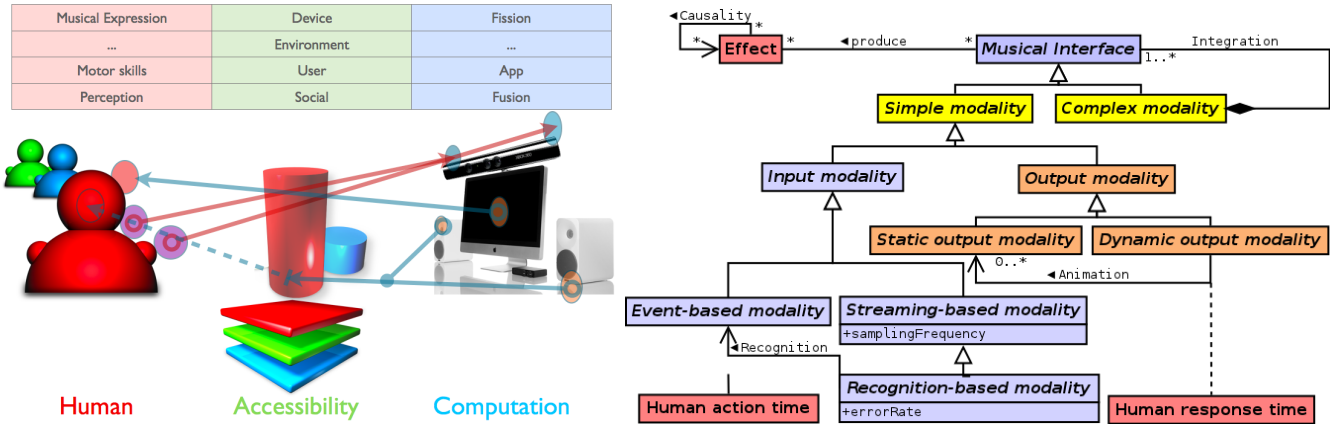

Figure 1: The structured design and evaluation model we are following, after [4] (left), and the corresponding multimodal interaction model (right). The human capabilities, computational modalities and accessibility issues are expressed as structured layers in this framework, where the latter is utilized to check how well the intentions of design are perceived by the user.

For instance, we can perform a percussive event recognition and classification on an audio stream, where the mainstream algorithms yield an error rate of 20 %. In computing and rendering the *output modalities*, we always consider the human response and its time-scale.

The model specializes the output modality as static or dynamic, with an *animation* association between the modalities, which is constrained by the *human response time-scale*. The human interactive response is considered at three levels: perceptual processing (about 0.1 second), immediate response (about 1 second), and unit task (about 10 seconds). For instance, the animation of static images is considered to produce a movie with smooth motion, if the duration of each image is less then the perceptual processing response time. For rhythmic interaction, the perceptual processing time of a *smear window* is important to perceive the event order, as a necessity of cognitive function [1].

2.2 Evaluation model

Similarly, the evaluation constraints can also be structurally decomposed in basic and complex constraints, and two main types of basic constraints can be identified: user and external constraints. The user constraints are user feature, user state (emotional and cognitive contexts), and user preference, whereas the external constraints are structured as device constraint, environmental constraint, and social context. The observations, remarks, and the evaluation outcomes then can be tabulated, similar to Fig. 1, left. Not all aspects may be evaluated in a single session, but they should still be kept in mind when designing the tests, and inference should be sought from the test results.

3. RHYTHMIC INTERACTION IN IPALMAS

iPalmas was developed for observing the rhythmic interaction of people with a maximally simple interactive system, to teach a novice user Flamenco hand clapping patterns [2]. We expect this type of interaction to engage people, without requiring any special skills. In the following, we elaborate the relation between the modeling framework presented in Sec. 2 and the design and evaluation of iPalmas.

3.1 iPalmas design model

iPalmas is designed for interaction between the user and a virtual tutor. The primary *input modality* is an audio stream of the user's performance. In producing this stream,

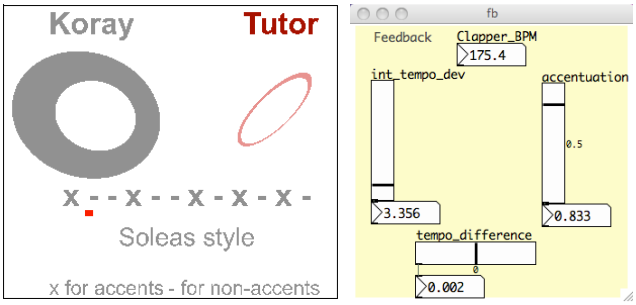

Figure 2: Three different visual displays of iPalmas. (Left) The circles to indicate the match between the tutor and the performer and the compas presentation. (Right) Various metrics presented by sliders and number boxes.

the user claps her hands in alignment with the virtual tutor, coordinates her motor action, and experiences an audio-tactile feedback naturally occuring during the clapping. As we know from sensorimotor synchronization studies, this multimodal feedback has implications on the timing of motor action [1].

The *streaming-based modality* of the user's clapping is converted to an *event-based modality* by real-time hand clap sound *recognition*. Each event is characterized by the event time, the detected hand configuration (cupped vs. straight hands), the detected accentuation (loud vs. soft claps), an update to the clapping tempo estimate, and an update to the estimate of temporal deviation in the clapping. While tempo and temporal deviation are continuous measures, they are updated only for each detected event. The system, as a *musical interface*, aims at coordinating this input modality with a *complex output modality* to present the user a Flamenco pattern to practice and an online evaluation of the user's learning and performance.

To achieve this, iPalmas utilizes both the *auditory* and *visual output modalities*. The target hand clap pattern is presented by synthetic hand clapping sounds within a reverberant environment (both *dynamic output modalities*), and a visual transcription of the accentuation (see Fig. 2, left). This transcription consists of 12 marks corresponding to the beats of a Flamenco compas (highlighted by shape, either - or x, depending on the compas), a red visual marker (highlighted by shape and color), the pattern name and the

legend of accents, all static output modalities. The visual display contains also the following four static output elements (highlighted by color): two circles and two textual elements indicating the name of the current user and the tutor. The color elements are also used for association of the circle with the text label. Here, the *perceptual effects* of *grouping by color and proximity* are in operation.

The visual *dynamic output modalities* include the animation of the visual marker and the two circles. The visual marker is animated by the tempo of the tutor, to indicate the current position within the pattern. This marker wraps to the beginning after the last beat of the compas, and a short auditory marker is played at the wrap-around. The same tempo animates the tutor circle (the right one), resetting its sway clockwise at each clap occurence. The user circle, on the left, is animated in a similar fashion, but the resets happen at each detected clap. The distance between the circles' centers gets smaller, when the user's clapping tempo gets closer to that of the tutor. At this step, we were aiming for a perceptual grouping both on *proximity* and *common fate gestalt*. The thickness of the circles indicates accentuation, with the thick circle corresponding to an accentuated clap. Finally, the circles sway clockwise and back for each (detected) clap, so when the user perfectly matches the tutor's performance, the circles move unanimously.

In addition to the abstract representation of the circles, the user is presented numeric metrics on the performance, indicating the difference between the user's and the tutor's tempo, the user's internal tempo deviation, and the incorrectness of performing the accentuation (the bottom part of Fig. 2). With perfect performance, all the metrics are zero. By using the GUI elements such as label texts, slider, and number boxes, we were aiming for *cognitive effects*.

3.2 iPalmas evaluation model

We have performed an evaluation of the iPalmas system with 16 subjects [2]. This number provided a good balance between the combinations needed by the experiment design and the discoverability of most of the usability problems with a small subject group (i.e., Nielsen's model, see [7]), in our design phase iteration.

Most of the participants had musical background, but none of them were Flamenco practitioners. They practiced four different hand clapping patterns, two with the auditory output only (hand clapping of the tutor) and two with both auditory and visual output (hand clapping, transcription, circles, and numeric metrics). In half of the cases the virtual tutor's tempo remained constant, in the rest the tempo was allowed a small drift from the original tempo, adapting to the user's clapping. In the experiment, the subjects first practiced a pattern and then performed the learned pattern for one minute without the tutor's hand clapping. The evaluation results are presented in Table 1, according to the evaluation model presented in Sec. 2.2. Since the evaluation was carried out in laboratory conditions with one subject at a time, the social and environment constraints were not tested. However, qualitative observations gathered from questionnaire and follow-up discussions provide some insights in these aspects. In the following, only the most important observations, indicated by Roman numbers in the table, will be discussed. The reader is referred to [2] for a more detailed discussion.

The auditory output was found to be the most important factor in learning the patterns (I). Out of the visual elements, the most useful one was the transcription of the pattern, with the moving marker below it (IIa,b). The rhythmic performance of the subjects varied between different pattern-tutor combinations and subjects, but in general it was found that the subjects tended to accelerate, once the auditory output faded away (III).

Some subjects showed more variation in their temporal performance than others (IV). Visual elements, namely the transcription (Va), and the allowed tempo adaptation (Vb), helped in succeeding with the accentuation. With a tempo-adaptive tutor, the time between two claps was slightly longer before an accentuated clap than before an un-accentuated clap. In general, the subjects regarded the numeric metrics (VI) and the dancing circles (VII) of limited use in the interaction and learning.

4. CURRENT DESIGN PHASE

The evaluation provided us good insights about our target group. We have considered the *user preferences* that have assessed the usefulness of auditory and visual markers, various visual elements, and especially the transcription, resulting in a new visual display, reported in the next subsection. In addition, the advanced auditory perception capabilities of some participants, who reported excessive reverberation and were disturbed by early reflections, inspired us to rely on the reverberation as an auditory display. Shortly, we are currently focusing on the audio-visual "touch-points" of iPalmas, and plan to revisit the technical aspects (device, application, system) in the next phase, finally completing the development cycle. The final design of iPalmas will be demonstrated at `http://www.acoustics.hut.fi/research/ipalmas.html`

4.1 Visual display

As observed in evaluation, having three separate graphical representations (clap pattern, metrics, and circles) did not resonate well with our subjects. A new graphical interface that unifies those three regions is under development. The concept is illustrated on Fig. 3. It is an abstraction of the traditional Flamenco compas. Note that the figure overlays several instances of visualization for brevity. The concept consists of twelve discs, arranged in a circular manner according to chosen clap pattern (Soleás in the figure). The numbers are optional, but included to stimulate the referential learning of rhythms by counting. The progress of time is represented both continuously (by a "fluid" flowing in the central, circular grey tube), and also discretely, by highlighting the position of the tutor. This highlighting can be done in several ways, either by a glow as presented in the figure on beat 3, or by simply hollowing out/refilling the particular circle, integrated with the tutor's clap.

The user activity is represented by rings, as the blue accented clap on beat 6, or the orange non-accented pattern on beat 9. When the tutor disc is highlighted simultaneously with the correct accent of the user, then a good performance is achieved. The performance indicators presented in Fig. 2 may also be used to modulate the radius of the central grey tube, in a way that the tube becomes infinitesimal when perfect performance is achieved (i.e., the performer does not need this performance measure anymore).

4.2 Auditory display and sonification

For a tight multimodal integration with the visual display, we also plan to sonify the key parameters of the user's performance. To start with, we have several parameters computed by the iPalmas system. These parameters can be divided into event-based and continuous sonification targets. The event-based targets include the correctness of each accent and the temporal offset from each of the tutor's claps. The continuous targets are tempo lead or lag (how much the user is clapping ahead/behind of the target tempo), overall

Table 1: The evaluation results of iPalmas, tabulated according to the model presented in Sec. 2.2. The social and environmental aspects, indicated by (*) are not directly observed.

	Social(*)	User	Environment(*)	Device, App
Sensory		Auditory marker distracting Visual marker distracting	Noise, masking	Cross-talk
Perceptual (Auditory)	No solo claps heard	Perceived "castanets" Reported excess reverb	Reverb	Synthetic sound Reverb algorithm
Perceptual (Visual)		Prefer static compas Audio-visual sync?		 Sync of threads
Motor	No solo claps practiced	Cannot produce accents **III. Speed up when tutor stops** Fatigue **IV. Temporal variation** **Va. Transcription helps accentuation**	Reverb	 Latency Latency
Cognitive		**I. Prefer Audio Feedback** Too many visual elements	Smear windows	Latency Threads
Memory	Comparison	2 subjects remembered all 4 patterns On average, 2 patterns remembered **IIa. Transcription helps recall**	 "Whole plus detail"	Pattern dictionary
Learning	Comparison	Prelistening (about 40 seconds) **VI. Metrics of limited use (for some)** **VII. Circles attractive, but not useful** **IIb. Transcription helps learning**	Noise, masking	Test phase design
Expression	Shared mastery	**Vb. Adaptive mode improves**	Smear windows	Critical latency

Figure 3: Concept for iPalmas visualization.

correctness in accentuation (computed with a running correctness metric), and the user's internal tempo deviation. The metrics were previously presented in numeric form, as in Fig. 2. In the sonification, we plan to concentrate on the continuous targets, as we consider them more important in continuous interaction.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a structured model, previously proposed for modeling multimodal interaction and evaluation, for the design and evaluation of musical interfaces. It can work as a tool for studying the existing design and evaluation cases, or can be used for informing the development. Here, we have utilized both on our own development, design, and evaluation of iPalmas.

While the structural decomposition of musical interfaces, interaction paradigms, and novel applications to atomic components may seem a difficult task at a first sight, we aim to build a collection of models and successful patterns [3]. Our other future task is to work out the evaluation results presented in Table 1 and complete the evaluation of social and environmental aspects. The *user features* we have observed can be summarized in a few user profiles, which may inform the next development phase. For instance, new training modes can be developed for the users who have never heard or practiced their clapping in isolation, but only in crowded concerts of similar social gatherings. On Table 1, we have correlated some user preferences with technical system components. Among them, timing, latency, and threads are crucial factors that we need to consider. Finally, we plan to evaluate the visual and auditory displays proposed in this work.

6. ACKNOWLEDGMENTS

This work is supported by the Academy of Finland (Pr. 140826), the Graduate School of Aalto ELEC, and the Aalto Media Factory. We acknowledge the work of Inger Ekman and Koray Tahiroğlu in the first evaluation and visualization of iPalmas, respectively. We also thank Antti Ikonen and Ferhat Şen for new visualization ideas and the evaluation participants for giving us a hand (or two) in exploring rhythmic interaction.

7. REFERENCES

[1] C. Chafe and M. Gurevich. Network time delay and ensemble accuracy: Effects of latency, asymmetry. *Proceedings of the AES 117th Convention*, Oct. 2004.

[2] A. Jylhä, I. Ekman, C. Erkut, and K. Tahiroglu. Design and evaluation of human-computer rhythmic interaction in a tutoring system. *Computer Music J.*, 2011. Accepted for publication.

[3] C. Larman. *Applying UML and Patterns : An Introduction to Object-Oriented Analysis and Design and Iterative Development (3rd Edition)*. Prentice Hall PTR, October 2004.

[4] Z. Obrenovic, J. Abascal, and D. Starcevic. Universal accessibility as a multimodal design issue. *Communications of the ACM*, 50(5):83–88, 2007.

[5] Z. Obrenovic and D. Starcevic. Modeling multimodal human-computer interaction. *IEEE Computer*, 37(9):65–72, 2004.

[6] S. O'Modhrain. A Framework for the Evaluation of Digital Musical Instruments. *Computer Music J.*, 35(1):28–42, 2011.

[7] A. Sears and J. Jacko. *The Human-Computer Interaction Handbook*. Fundamentals, evolving technologies, and emerging applications. Lawrence Erlbaum Associates, 2nd edition, 2008.