EmoSonics – Interactive Sound Interfaces for the Externalization of Emotions

Thomas Hermann
Ambient Intelligence Group
CITEC, Bielefeld University
Bielefeld, Germany
thermann@techfak.uni-bielefeld.de

Jiajun Yang
Ambient Intelligence Group
CITEC, Bielefeld University
Bielefeld, Germany
jyang@techfak.uni-bielefeld.de

Yukie Nagai
Graduate School of Engineering, Osaka University
Osaka, Japan
yukie@ams.eng.osaka-u.ac.jp

ABSTRACT

This paper presents a novel approach for using sound to *externalize emotional states* so that they become an object for communication and reflection both for the users themselves and for interaction with other users such as peers, parents or therapists. We present an abstract, vocal, and physiology-based sound synthesis model whose sound space each covers various emotional associations. The key idea in our approach is to use an *evolutionary optimization* approach to enable users to find emotional prototypes which are then in turn fed into a *kernel-regression-based mapping* to allow users to navigate the sound space via a low-dimensional interface, which can be controlled in a playful way via tablet interactions. The method is intended to be used for supporting people with autism spectrum disorder.

CCS Concepts

•**Human-centered computing** → **Auditory feedback;** *Accessibility technologies; Accessibility systems and tools;*
•**Applied computing** → *Sound and music computing;*

Keywords

Emotions, Sound, Auditory Display, Autism Spectrum Disorder (ASD)

1. INTRODUCTION

Emotions play an important role for our experience of the world and ourselves as they color experience and influence our decisions and their execution. Furthermore, the expression of emotions and perception of emotions are highly relevant for social interaction. As emotional intelligence is a complex function of the mind, and it operates usually automatically we are rarely aware of how exactly the dynamics of emotions operates and manifests. Emotions manifest normally in a number of carriers such as facial expression, movement, prosody in spoken language, or explicitly by the choice of spoken words. They furthermore correlate and influence physiological states such as heart rate, body tension, etc. Yet if – for instance due to certain dysfunctions or disorders such as Autism Spectrum Disorder (ASD) – the perception, processing and manifestation of emotional signals is hindered, it affects the individual and social live significantly [9]. Failing to express feelings or pain may lead to many negative effects including increasing in anger or even self-injuries [12]. This is risky as people suffering from ASD are often not able to seek help using methods such as facial expression. Subsequently, caregivers are not able to spot it for the same reason. A study shows that a person suffering from ASD may appear to be completely calm while having an unusually high resting heart rate [4].

With this research we set out to develop a *sound-based interface* which might to some degree bypass cognitive processing steps and facilitate the expression of emotions in a direct way. Using sound is motivated from the fact that it is already established and known to be an important carrier of emotional charge, for instance consider film music, the expressiveness achieved by prosody, or music in general. By means of the interactive synthesis of emotional sounds, and an iterative refinement of the sounds to match the innerly felt emotions, the user is expected to perform an inner 'self-reflective dialogue' for which the synthesized sound becomes a mirror image of the sensed state. This may help users (a) for themselves to pay more attention to their emotions, understand and observe them more clearly, (b) to simultaneously find novel ways to express them, i.e. to bring them beyond the surface, which is particularly relevant if emotional production is hindered, and (c) to render emotions more 'tangible' as a shared resource to be worked with, for instance, as a therapeutical means.

ASD patients could profit from such a technique that allows them to express emotions towards others (peers, caregivers, parents), or even to train the perception of emotional categories as they manifest in sound in gamelike interactions.

In this paper we will first review related work at the intersection of sound and emotion, then outline basic assumptions about and representations of emotions as a basis for the definition of *continuous* sound models that enable the expression of various emotional signals and their interpolation and morphing. These models necessarily have a large number of parameters, which complicates the adjustment towards clear emotional expressions. Therefore we proceed with an approach inspired by evolutionary optimization techniques where the user merely iteratively selects one of

AM '16, October 04-06, 2016, Norrköping, Sweden

DOI: http://dx.doi.org/10.1145/2986416.2986437

several sound variations in search of a sound that matches a given emotional prototype. This allows us to investigate the clustering (e.g. dispersion, homogeneity) of parameter vectors in sound model space for certain primary emotions, and subsequent definition of anchor points for kernel regression. In the following section we apply kernel-regression to provide a mapping between an interaction space (low-dimensional manifold) and the high-dimensional parameter space of sound models by means of which we can reduce the complexity of specifying and interpolating sounds that represent emotions significantly. We illustrate this by various sound examples. As the focus is on the technique, we only touch on some prospects for evaluation of the new approach before ending with discussion and conclusion.

2. RELATED WORK

Currently, the studies of sound and emotion mainly focus on three domains: music, speech and acoustic events.

The conventional way of measuring the emotional response from listening to music is to ask subjects to verbally express their experience, or using eclectic scales method to weight different emotional labels [15]. Koelsch et al. [10] used functional magnetic resonance imaging (fMRI) to study how consonant and dissonant music triggered different brain activities corresponding to different emotions.

Film music is considered by Schubert as a sonification process of conveying emotion that guides film viewers to the intended emotional context [17]. In [17], Schubert et al. concluded the difficulty of sonifying emotion via conducting a study based on listening to pieces of different film music. The study showed that it is easier to quantify the low-level emotional arousal via parameter mapping of acoustic parameters such as sound pressure level (SPL), pitch or tempo. Yet when lending to complex emotions it became increasingly harder and film music often serves as a complex entity to deliver guidance to the emotion but not without ambiguity. Winters and Wanderley explored continuous sonification [6] to monitor arousal and valence [20]. They looked into low-level parameters such as fundamental frequency or loudness, in contrast to complex musical theories.

A wide range of studies have been conducted in the field of speech synthesis to simulate emotion and making speech synthesis sounding more natural. In [16], Schröder reviewed various studies on formant and time-domain synthesis and provided a parametric guideline. Tachibana et al. created a HMM-based speech synthesis allowing interpolation between two different emotional speaking styles [18]. Such research mainly focuses on robotic voice enhancements such as prosodic text-to-speech systems or how to make robots sound more human-like.

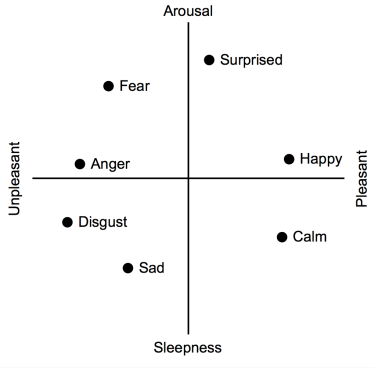

Figure 1: The circumplex model of emotions (modified from [14]).

In comparison with the presented examples, our approach aims at providing both expressive and complex models which are capable of rendering audio based on a target emotional state and at the same time also of enabling seamless interpolation between the sounds for different emotional states. Although the control space of the sound models is high-dimensional, the kernel-regression algorithm enables its navigation from a rather low-dimensional interaction space. This, in result, allows the navigation of emotional sounds to be realized even on a small-scale 2D display of small devices such as smartphone or tablet computer.

3. REPRESENTATION OF EMOTIONS

Emotions are perceived based on physiological and/or contextual states. It is suggested that people have six basic emotions (happiness, surprise, fear, anger, disgust and sadness) regardless of their cultural background. Traditional theories suggested that the basic emotions act as discrete categories because they are distinguishable from each other by their physiological responses and behavioral expressions.

In contrast, Russell [14] proposed a continuous representation of emotions. His model, referred as a circumplex model, defines emotions in a two-dimensional space consisting of pleasant/unpleasant and arousal/sleepiness axes. Fig. 1 illustrates the model, in which the six basic emotions plus calm are plotted in the 2D space. For example, happiness is perceived as a state with high pleasure and a medium level of arousal, whereas surprised and sad feelings are characterized by high and low arousal respectively. An important point is that emotions can be represented at any level of pleasure and arousal. Common neurophysiological systems are suggested to produce and perceive all possible emotional states [13]. Our approach to sonically externalize emotional states is inspired by such model.

4. DESIGN OF SOUND MODELS FOR EMOTION COMMUNICATION

The International Affective Digitized Sounds (IADS) provides a standardized sound library and the association with the different emotional charge. It exemplifies that sounds cover a large spectrum of emotional associations in the pleasure–arousal–dominance space [2]. The problem with their used sounds, however, is their heterogeneity, i.e. the differences between sound sources, stretching from environmental nature sounds to violent inter-human interactions and, as side-effect, the differences in emotional associations that are likely triggered. For instance, the affective charge may strongly depend on the cultural background and experience of the listeners. In another cross-cultural study from [11], the emotional expressions of sounds (such as loud, strong, noisy, pretty) differed between subjects from five different countries given the same collection of sounds. For that reason we conclude that the most important requirement for our target sound models should be

R1 expressiveness: the sounds should be able to express (resp. evoke) a rich variety of emotions, best cover-

ing the complete emotional space, i.e. the whole pleasure/arousal space.

R2 continuity: the sounds should be the result of a sound synthesis where any (small) parameter variation continuously results in a (small) change of the sound, thus allowing to meaningfully interpolate between parameter vectors.

R3 culture dependency: the sounds should have a low cultural dependency. This minimizes misinterpretation of sounds subject to the user's cultural background. To achieve this, we consider artificial and synthesized sounds that do not associate strongly to specific cultural knowledge. The design also excludes embodiment of musical characteristics.

In the course of developing sound models to express emotions we discovered that stand-alone events allow to identify an affective charge rather quickly, within few seconds. These events could be regarded as cartoon versions of emotions. Similar to cartoon figures which often communicate an emotional state very concisely and sparsely, just by using few pencil strokes, the sound events can be regarded as cartoonified emotion sounds. Such short sounds are particularly attractive for tightly-closed interaction loops in which users refine sounds, since it shortens the cycle duration. For that reason we add as secondary requirement:

R4 The sound should convey the emotion within a short time frame of less than 2.5 seconds.

Furthermore, we acknowledge the problem of controllability, particularly for highly expressive yet non-parametric sound models. For example, an additive synthesis model using 50 harmonics and 20 control points over time would result in 1000 parameters. Since, however, we are interested in manually still controllable sound models, we add as further constraint:

R5 The number of parameters should be low, i.e. smaller than ≈20, to enable us in principle to also practically adjust the parameters manually.

In search of models to meet R1–R5, our subjectively biased design attempts converged to three models which we describe next. These are abstract sound model, vocal sound model and physiological sound simulation. As there are many other sound design possibilities, for the current stage, these three models covers a suitable range of sound creation range from the most raw approach (abstract sound model) to mimicking the familiar way of expressing emotions (vocalisation thus the vocal sound model), and even using sound to represent our emotions when certain emotion is appeared.

4.1 Abstract Sound Model

We started with the idea of a model that uses very abstract sounds – sounds that cannot be as easily interpreted as coming from any natural cause and thus comply with R5.

The model is implemented as a synth on Max 7.2.1 and provided with 11 controllable parameters. The synth creates a baseline timbre based on a combination of three common wave forms: sine, square and sawtooth. The synthesis is triggered by events meaning that at each trigger the sound appears with certain duration rather than being rendered

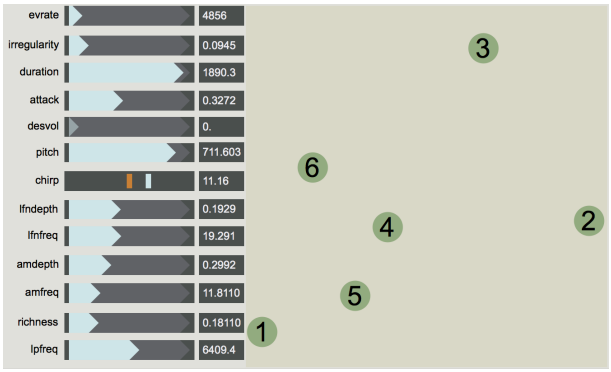

Figure 2: Slider- and Node- based GUI for parameter adjustment of the abstract sound synthesizer.

parameter	range and unit
duration	[0.01, 2.0] s
attack	[1, 80] %
decay slope	$[-40, 0]$ dB/rm time
AM intensity	[0, 1]
AM rate	[0, 60] Hz
pitch	[20, 85] midinote
chirp	$[-36, 36]$ semitones
LFO rate	[0, 50] Hz
LFO intensity	[0, 1] rel. of pitch
spectral richness	[0, 1] mix between sin, triangle, saw
LP cutoff	$[2000, 10000]$ Hz

Table 1: Parameters and ranges for the abstract synthesis sound model.

continuously. Within the duration of the sonic event, carefully selected modulations and variations are applied. The parameter GUI is shown in Fig. 2. First of all, the `duration` parameter d decides how long each segment of sound is, setting $d \in [0.01, 2.0]s$. The `attack rate` is a linear volume ramp that decides how much percentage of the beginning signal is going to be smoothly ramped up from 0, resulting in a range between percussive and slowly appearing sounds. An initial `pitch` $p \in [20, 85]$ (MIDI scale which translated to 25 1109Hz) can be selected, then the `chirp` parameter determines the amount in semitones (±3 octaves) that the ending pitch is changed relative to the initial pitch. A low frequency oscillator (LFO) is added to the carrier as frequency modulation ($\mathrm{FM}_{\mathrm{freq}} \in [0, 50]$ Hz, $\mathrm{FM}_{\mathrm{int}} \in [0, 1]$), adding a little more natural feeling to sound when the modulation amount is very small. In its extreme, the vibrato effect intensifies and the sound becomes more rough. The amplitude modulation ($\mathrm{AM}_{\mathrm{freq}} \in [0, 60]$ Hz, $\mathrm{AM}_{\mathrm{int}} \in [0, 1]$) adds a tremolo effect to the tone, i.g. similar to the trembling in the voice if a person is nervous. Finally, the `richness` parameter alters the mix between the above mentioned waveforms and `lpfreq` controls the cut-off frequency of a lowpass filter that shapes the brightness of the sound. The parameter range is shown in Table 1.

4.2 Vocal Sound Model

This sound model is inspired by the human voice as formidable carrier for emotional charge, even without articulation of meaningful words, i.e. without using language. As

```
SynthDef(\vs, {| out=0, amp=1, pitch=50, chirp=0, dur=0.5, att=0.0, decslope=(-12),
  amint=0, amfreq=0, lfnfrq=0, lfnint=0, vowel=2, voweldiff=0, bright=1, pan=0 |
var sig, aenv, fenv, amsig, va, ve, vi, vo, vu, blend, r=0.01;
amsig = SinOsc.kr(amfreq, mul: 0.5*amint, add: 0.5);
aenv = Line.ar(0, decslope,dur, doneAction: 2).dbamp *
EnvGen.kr(Env.new([0,1,1,0], [att*dur, dur*(1-att)-r, r]));
fenv = Line.kr(pitch, pitch+chirp, dur).midicps + LFNoise1.kr(lfnfrq,lfnint*pitch.midicps);
vu = Vowel(\u, \tenor); vo = Vowel(\o, \tenor); va = Vowel(\a, \tenor);
ve = Vowel(\e, \tenor); vi = Vowel(\i, \tenor);
blend = Line.kr(vowel, vowel+voweldiff, dur);
sig =  Formants.ar(fenv, vu
    .blend(vo, blend.linlin(0,1,0,1,\minmax)).blend(va, blend.linlin(1,2,0,1,\minmax))
    .blend(ve, blend.linlin(2,3,0,1,\minmax)).blend(vi, blend.linlin(3,4,0,1,\minmax))
    .brightenExp(bright.reciprocal, 1));
Out.ar(0, Pan2.ar(sig * amsig * aenv;, pan, amp));
}).add();
```

Figure 3: SuperCollider code for the vocal synth definition.

the human voice can be simply modeled via a source/filter model we here apply subtractive synthesis with controllable formants to define the synths. The parameters for the models are kept largely in analogy with the previously described abstract sound model. Specifically, we implement the duration, amplitude envelope, amplitude modulation, pitch, chirp, and frequency modulation parameters in the same way. For practical implementation we here used SuperCollider 3.6, using the Vowel class described in Grond et al. [5] to implement the formant/filtering part of the subtractive model.

Two formants suffice for the perception of vowels. However, to gain better discernible sounds we use 5 formants, and specifically the tenor settings for a male voice in the middle pitch range. Since, however, 5 formants would require alone 15 parameters (5 center frequencies, band widths and intensities), a full-fledged control would violate R4. Hence we reduced the timbre control to only one parameter: an interpolation coefficient $v \in [0, 4]$ over the vowel range "u-o-a-e-i" (vowels roughly pronounced as in 'moon-bow-bar-bed-in'), so that $v = 0, 1, 2, 3, 4$ results in 'u', 'o', 'a', 'e', 'i' and real values in between interpolate linearly between the vowels. The order of vowels was chosen to implement a linear order from closed&dark vowels to open&bright vowels.

A constant vowel, however, gives a limited perception in terms of vocal contour. For that reason we introduce a vocal drift parameter $v_{\text{drift}} \in [-2.5, 2.5]$ by means of which the vowel at the end of the event becomes $v + v_{\text{drift}}$. On top of these two parameters we use a brightness parameter to control the spectral richness of the sound, ranging from rather mellow sounds to bright sounds with lots of harmonics. The vocal parameters are demonstrated in interaction video V_1 [1]. Table 2 depicts the parameters of the vocal synthesizer.

In combination with the other parameter the resulting sounds can (in certain parameter ranges) be associated to utterances or vocalizations.

4.3 Physiology-inspired Sound Model

There is strong evidence that emotions affect the physiological responses of the body. Arousal as an aspect of emotion, for instance, manifests also in aroused physiological responses such as heart rate, respiration, or skin temperature, muscle spasms, fatigue, constriction in internal organs, etc. [3]. The physiologically inspired sound model starts from the assumption that a synthesis of bodily sounds that correspond to originating emotional causes can elicit or induce in the listener a similar emotion as remembered from own experience.

parameter	range and unit	map	D_0
duration	[0.005, 1.5] s	exp	0.4
attack time	[0.001, 0.5] s	exp	0.001
decay slope	$[-50, 10]$ dB/rm time	lin	-12
AM intensity	[0, 1]	lin	0
AM rate	[1, 50] Hz	exp	1
pitch	[20, 85] midinote	lin	50
chirp	$[-36, 36]$ semitones	lin	0
LFO rate	[5, 50] Hz	exp	5
LFO intensity	[0, 0.5] rel. of pitch	lin	0
vowel	[0, 4] "uoaei"	lin	2.5
vowel drift	$[-2.5, 2.5]$	lin	0
brightness	[0.2, 1]	lin	0.5

Table 2: Parameters and ranges for the vocal synthesis sound model. The column 'map' states whether this parameter is linearly (lin) or exponentially (exp) mapped from the interval [0,1].

Some of these bodily manifestations already possess an audible expression, such as heartbeat and breathing sounds, which are selected to be synthesized in this study. The rationale to implement audio synthesis techniques rather than using recorded samples is so that they offer much more freedom of parametrization. Furthermore, users can navigate the parameter space seamlessly via interpolation and morphing. Using recorded samples, on the other hand is much more rigid.

In extension beyond the directly audible, however, we consider physiological reactions that are not audible yet that may be nonetheless epressed by synthesized sound. Trembling, muscle tension and blinking of the eyes are examples for such non-audible phenomena. We postpone these physiological components for future explorations. The following section explains the two physiological sounds in details.

4.3.1 Synthesizing Heartbeat sounds

Heartbeat can normally be heard via a stethoscope. The sound of a heartbeat comes from a two-part pumping action of the heart, which usually takes about one second. Our heartbeat synthesizer is based on subtractive synthesis and produces a periodic sound. It uses three triangular waves (f_0, $2f_0$ & $0.5f_0$). This provides a rich harmonic content for the filtering. The carrier frequency ranges between 20 and $100Hz$. The signal passes through an amplitude envelope that creates the temporal amplitude pattern of the heartbeat. For that, the envelope creates two percussive hits per cycle. Finally the signal passes through an LFO-controlled lowpass filter, where the cutoff frequency is modulated to make the sound closer to a real heartbeat sound. With increasing LFO frequency the sound changes from smooth to rough. A synthesis structure is presented in Fig. 4.

The parameters are listed in Table 3: `pitch` defines the fundamental frequency of the sound.

While a lower frequency (< 50 Hz) will sound more natural in general, a higher pitch makes the sound easier to be perceived. The rhythm decides the repetition rate of the beats. The human heart rate is typically between 60 and 200 beats per minute (bpm). Typically, a slower rate is often associated with a calm and lower arousal state, such as sleep (40 – 50 bpm). On the other side, a higher heart rate is connected with high arousal states which can be both

[1] Media files are provided at http://doi.org/10.4119/unibi/2905039

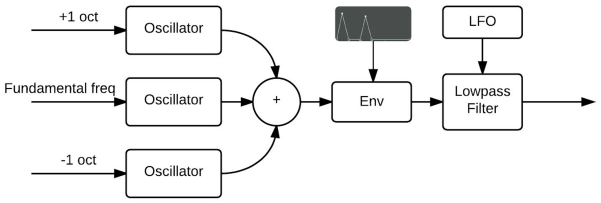

Figure 4: Heartbeat synthesis diagram.

parameter	range/unit	physical correspondent
pitch	[20, 240] Hz	for resulting lower but perceivable tone.
rhythm	[30, 200] bpm	speed of the heartbeat.
LFO rate	[0, 40] Hz	intensity of the heartbeat.

Table 3: Parameters and their correspondent physiological ranges for the heartbeat synthesis.

positive (excited) and negative (angry & feared). As last control parameter, the `LFO rate` alters the intensity of the beat itself ranging from clear and smooth to rough and distorted, which can be used to distinguish the pleasantness of the emotion.

4.3.2 *Synthesizing Breathing Sounds*

Respiration is the inhalation and exhalation of air. The process facilitates several of our key organs. The air passes through mouth or nose and is then sucked in and passed out of the lungs due to the expansion and contraction of the thorax. The sound of respiration is made audible because of the air passing through our nose or mouth. It is generally not a pitched sound but noise-like. Breathing through the nose has more high frequency content ('hiss') while breathing through the mouth is spectrally lower and perceptually more thick ('huu'). Thus, we use a pink noise generator as the sound source and apply a comb filter for fine tuning (see Fig. 5).

To separate nasal breathing and open-mouth breathing, the filtered noise is passed through two extra filters: a low-pass filter ($f_c \approx 150\,\text{Hz}$) and a high-pass filter with cutoff frequency ($f_c \approx 1000\,\text{Hz}$). These separate signals are mixed at different amplitude levels, which controls the balance between the two breathing types.

The action of the breathing is split into three sections: breath in, mid gap and breath out. For the inhalation and exhalation, two multi-point graphical envelopes are used allowing fine-tuning of the volume change in such actions. The up-ramping envelope is for simulating inhalation with a duration of $t_{\text{in}} \in [0.2, 2]\,s$. The down-ramping (exhalation) time is relative to t_{in}. $t_{\text{gap}} \in [0, 0.5]\,s$ is the time gap between the two phases. Interestingly, just by varying the relationship between the three time variables, a wide range of breathing types can be simulated. As the last step, the rhythm parameters determine the rate of repetition. The list of controllable parameters are shown in Table 4.

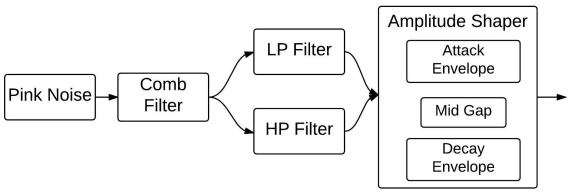

Figure 5: Breathing synthesis diagram.

parameter	range/unit	physical correspondent
t_{in}	[0.2, 2] s	the breath in time.
t_{out}	[0, 5] rel. of t_{in}	relative of the breath in time.
t_{gap}	[0, 0.5] s	gap between the two phases.
mix	[0, 1]	balance between mouth and nasal breathing.
rhythm	[30, 200] bpm	the rate of breathing

Table 4: Parameters and their correspondent physiological changes of the breathing synthesis.

5. EVOLUTIONARY OPTIMIZATION OF EMOTIONAL PROTOTYPES

Given that the above mentioned sound models fulfill our requirements and are expressive in terms of emotional quality, it is now interesting to understand (a) how far users agree upon emotional assignments to sounds resulting from a specific parameter vector, i.e. the accuracy of sound-to-emotion mapping, and (b) what parameter space regions are chosen to obtain sounds that represent a certain emotion, i.e. the uniqueness of the emotion-to-sound mapping, or the dispersion of emotional sound categories.

In the context of auditory display, these two angles have been discussed in the study of auditory icons under the terms causal uncertainty and typicality [1]. In contrast to his discussion where sound sources were related to their evoked interpretation, we here relate emotional sounds to their induced emotional interpretation.

The problems can be studied in two different ways, of which we here focus on the emotional sound category clusters. Our empirical approach is to ask users to imagine a specific target emotion and then to refine the sound model so that the resulting sound suits to their option optimally. Sound designers are familiar with this task as it is akin the task of creating presets that sound like a pre-given musical instrument. In this vein, we ourselves first tried to adjust the parameters using a slider-based GUI as shown in Fig. 2. In this user interface the sound event occurs repeatedly at a user-adjustable rate and regularity, so that adjustments of the sliders deliver immediate changes of the sound. Such a parameter-aligned control can be effectively used by people who have some sound design experience, and is probably very hard to use for unskilled users. But for both groups we believe that this type of control is not the one that is most conducive for the discovery of the optimal emotional sounds, because it isolates parameters that ideally are meaningfully controlled together. Also, the GUI distracts from fully attending to the sound as such. For that reason we proposed and implemented a method that veils all parameter details from the participants and leaves to them solely the task to select one of 4 new candidate sounds which are derived from the starting sound (or to keep the previous sound). The

method is akin to evolutionary optimization and has been introduced and used by the authors before in [8] for the refinement of sonification mappings. The variations are obtained by modifying the initial parameter vector by means of random mutations. In evolutionary optimization, usually a fitness function is evaluated in order to select the descendants that survive for the next generation. In our approach, the fitness function-based selection is replaced by the manual selection. In consequence, the user enforces an evolutionary drift that lets the sound converge towards a fix point. Convergence is guaranteed by iteratively reducing the mutation strengths as detailed below. In summary, with few update steps, users – even without any sound design experience – are enabled to optimize sounds while being disburdened from any parameter-specific control GUIs.

5.1 Evolutionary Sound Design

Practically, we map the d-dimensional parameter space to the unit cube in $\mathbb{R}^d$ using a mapping function f^{-1}. The initial parent for the evolutionary process is thus

$$\vec{v}(t=0) = f^{-1}(\vec{p}) \ , \tag{1}$$

where $\vec{p}$ is the vector of default parameter values of the synths, as shown in Table 1, 2. Note that the mapping function f in its components is either linear or exponential, as defined in the table. The warping function serves to linearize the perceptual effect when making an additive change in the source representation.

From the parent $\vec{v}(t)$ in iteration t, n_c variations (called 'children') are rendered using a random process

$$\vec{v}_i(t+1) = \vec{v}(t) + \vec{\eta}(\sigma) \ \forall \ i \in [1, n_c] \tag{2}$$

where $\vec{\eta}$ is a vector of Gaussian distributed random numbers with mean 0 and variance σ^2. In each iteration, σ is decreased by multiplying with a factor $\alpha < 1$.

As user interface, as depicted in Fig. 6, on each iteration t, a radio button with $1 + n_c$ states is presented to the participant, where the first triggers playback of the parent $s(\vec{p}(t)) = s(f(\vec{v}(t)))$, followed by the sounds of the n_c children on the subsequent buttons. After listening to any of these sounds as often as wished the user clicks the 'proceed' button. This makes the parameters for the currently selected radio button the parent of the next generation, allowing also to reject all children by selecting the first button.

Figure 6: User Interface for evolutionary optimization of emotional sounds

In consequence, the initially large σ grants that the first iterations provide highly different sounds. With increasing generations, the smaller σ offers minor variations and thus rather subtle sound event adjustments. Whenever the user is sufficiently content, he or she can click on 'accept' to complete the episode. With the actual parameter settings, the six primary emotions are processed in $\approx$ 10–15 minutes.

5.2 Results: Sounds for primary emotions

We illustrate typical outcomes of the previously described method in sound example S1.1 – S1.6 (for 'happy', 'surprised', 'angry', 'disgusted', 'sad', 'calm'), which we also provide online on our website [2]

As you can hear in these sounds resulting from the vocal sound model, the sounds for happy, surprised, angry, disgusted, sad and calm share some specific properties and make sense. We can also hear variations within the groups, since different users conceptualize these emotions differently using the vocal synthesizer. Sound example S3 presents the sounds for the 6 primary emotions which result from averaging all parameters within the emotion cluster. It is remarkable that the parameter averaging produces sounds that actually capture relevant aspects of the individual sounds. This indicates that the idea of interpolation between sounds indeed works. The global mean of all parameter vectors of all emotions is a rather neutral sound, as can be heard in sound example S4.

Looking at the distance matrices between these cluster centers, as depicted in Fig. 7, we discover a neighborhood structure that is similar to the circumplex model: (a) happy and surprised are obviously nearby, and so are angry and disgusted, and to a lower degree sad. As emotions are sorted counterclockwise along the model, we would expect low values on the next-to-diagonal in the matrix. The most evident observation is the high distance between surprised and angry, though. The analysis was done using only 19 data points (8 for abstract, 11 for vocal). We plan to collect more data to investigate the clustering structure in more detail.

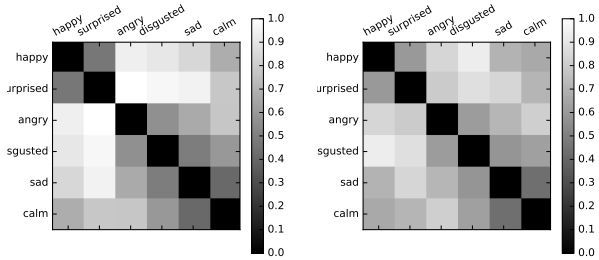

Figure 7: Distance matrices of cluster means for abstract (left) and vocal (right) sound model.

6. KERNEL-REGRESSION MAPPING IN EMOTION SPACES

This section presents an approach for mapping between a low-dimensional interaction space and the high-dimensional parameter space of the sound models. The need for such a mapping arises from the fact that most users are unable to cope with the challenge of directly controlling multi-parameter interfaces: it requires too much learning, too much time, and involves too much detail. Reducing the number of parameters by removing those that seem to be least relevant (i.e. those whose removal and replacement by average values would the least degrade the perceptual quality of the emotions) would provide some workaround, how-

[2]Media files are provided at http://doi.org/10.4119/unibi/2905039

ever, a reduction to as few as 1 or 2 parameters would result in sounds that are not anymore emotionally compelling. For that reason we propose the *kernel-regression mapping* to solve the problem in a way that gives us full control both over the interaction space manifold and the parameter space. This approach is very general, but for better understanding the idea, consider that we wish to navigate a 2D-space, e.g. a multitouch tablet surface depicting $N = 6$ emotion prototypes at locations $\vec{x}_i$. At each of these locations we'd like to reach a d-dimensional parameter vector $\vec{p}_i$. We need a controllable mapping that allows a continuous transition between an analogic and symbolic mapping.

This is exactly what Kernel regression mapping delivers, which has been introduced by the authors in [7] to better control a vocal mapping of EEG feature trajectories. Kernel regression is a standard approach to smoothly interpolate between a set of function values (outputs) defined at locations in input space [19], and we just quickly recapitulate the key idea before we show how it is applied here for the mapping task.

Kernel regression mapping calculates the parameter value at an arbitrary position $\vec{x}$ in input space (which here is our interaction space) by

$$p(\vec{x}) = \frac{\sum_{i=1}^{N} K_\sigma(\vec{x}, \vec{x}_i) \cdot p_i}{\sum_{i=1}^{N} K_\sigma(\vec{x}, \vec{x}_i)} \quad (3)$$

where $K(\vec{x}, \vec{y})$, is a kernel function that describes the relative importance of $\vec{y}$ to influence the interpolation at location $\vec{x}$. A standard choice for $K(\vec{x}, \vec{y})$ is the normal distribution $\mathcal{N}_{\mu,\sigma}(\vec{x} - \vec{y})$ with mean $\mu = 0$. The standard deviation σ plays a central role:

- if σ is small compared to the distances between prototype pairs in input space, the nearest input vector $\vec{x}_{nn}$ to a sample location $\vec{x}$ dominates the output, i.e. $p(\vec{x})$ deviates only slightly from $p(\vec{x}_{nn})$. In consequence, the input vector is tesselated into the so-called Voronoi cells, of strongly categorical character, i.e. we get a representation that promotes clearly separated and discernible prototype sounds and thus a *symbolic representation*. In the limiting case $\sigma = 0$ the parameters within the cell are constant within the Voronoi cells.

- with increasing σ, the parameter value will more and more be influenced by the neighboring prototypes and thus enable a meaningful continuous morphing of sounds between prototypes. A good value for σ is half the average over nearest prototype distances as seen from each prototype. We might call this an analogic representation, to contrast from the previous case. Apparently the kernel regression mapping helps overcoming the dichotomy of analogic/symbolic via a single control parameter.

- The limiting case, $\sigma \to \infty$, is of little interest as it would result in the mean $p(\vec{x}) = \sum_i p_i/N$ for any point in interaction space.

The following section will demonstrate how navigation on a 2D manifold allows straightforward navigation of emotional sounds.

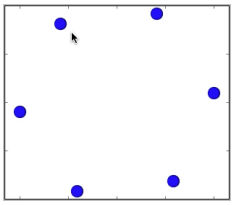

Figure 8: 2D interface, the nodes refer to emotion prototypes (clockwise from top: surprised, happy, calm, sad, disgust, angry) and are currently at fixed locations. Interpolated sounds can be interactively triggered via tapping (resp. Mouse click).

7. NAVIGATING EMOTION MANIFOLDS WITH A 2D INTERFACE

In this section we describe our first approach for a simple 2D interaction space. Given that our parameter space is around 15-dimensional this might appear an overly extreme reduction. However, we aim at very easy-to-use interactions on multitouch tablets, and for this 2D is the obvious choice.

The circumplex model suggests a ring topology for the primary emotions if organized along the pleasantness/arousal dimensions. From that observation we start by defining the prototype vectors $\vec{x}_i, i = 1...6$ for primary emotions i along the circle as depicted in Fig. 8.

Interaction video V5 shows how typical interaction yields sounds in the expectable fashion. Our subjective impression is that the interface is easy to learn and use. Note that a fixed spatial layout of the prototypes favors continuous interpolations between adjacent prototypes, which may appear as a limitation. For that reason we propose to allow the user to position (i.e. to drag and drop) the prototypes themselves in interaction space, allowing for instance interpolations between calm and angry prototype sounds by grouping them next to each other and others out of the way. Whether this flexibility is useful or even counterproductive depends on the target application, the users, and needs evaluation in studies.

We envision that autistic children might like to associate personally selected graphical icons with the prototypes, making the interface a more storyline experience. Maybe users might even want to create their own personalized prototypes and add them to the interaction space, or have a hierarchic approach, i.e. a number of prototypes just for a single primary emotion to better navigate within a cluster. We now have an environment that allows us to investigate such ideas without much effort.

8. DISCUSSION & CONCLUSIONS

We have demonstrated emotional sound models that have a moderately high-dimensional parameter space yet enable shaping complex *emotionally charged* sounds, i.e. they can represent the users' impression of certain emotional states. We introduced ways to *externalize emotions* using these sounds with our abstract, vocal and physiological sound models. We implemented an evolutionary optimization that simplifies exploration of the sound space by hiding the control parameters completely: Users simply select 'mutations' of the sound that points toward a certain 'emotional direction' and thus gradually refine and converge the sound into

a satisfactory emotion fingerprint. This process allows us to collect typical sounds that represent a selection of emotions from participants, agnostic to their audio synthesis background or skills.

We collected so far about 10 prototype vectors for each emotion, from different users both under the abstract and vocal model. We subjectively find that the sounds share some properties, i.e. there is a plausible intra-cluster coherence while at the same time the clusters exhibit distinct perceptual intra-cluster dissimilarity. At this time we lack sufficient data to analyze the clustering properties quantitatively, yet we aim to do that after conducting a user study with larger number of participants.

Our kernel regression mapping provides access to the high-dimensional data space from a low-dimensional control manifold, allowing a highly intuitive point & click exploration of relevant sound interpolations in a 2D interface. This provides the backbone of achieving an easy-to-use tablet-based *emotion externalization device* that can either function as a communicative tool, or as a game-like learning tool – especially for people who are suffering from ASD and having difficulties in expressing their feelings.

Therefore, the continuation of the current research will be looking into creating a tablet-based interface for sound sculpting. As future continuation we plan a study of letting ASD patients explore their own emotions via such an interface. It is hypothesized that ASD patients will yield different parametric region than typically developed persons. Hence, to achieve meaningful interpretation, we aim to develop a calibration system for individuals so that despite difference in emotional categories in correspondence to the parameter space, the generated sound can be 'translated' and be understood by others.

Acknowledgments

This research was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG); and by the EU FP7 project CODEFROR. The research has also been partially supported by MEXT Grant-in-Aid for Scientific Research on Innovative Areas (Research Project Number: JP24119003) and JSPS Grant-in-Aid for Specially Promoted Research (Research Project Numbers: JP24000012).

9. REFERENCES

[1] J. A. Ballas. Common factors in the identification of an assortment of brief everyday sounds. *Journal of experimental psychology: human perception and performance*, 19(2):250, 1993.

[2] M. M. Bradley and P. J. Lang. The international affective digitized sounds (2nd edition; iads-2): Affective ratings of sounds and instruction manual. Technical report, University of Florida, 2007.

[3] P. Ekman. Expression and the nature of emotion. *Approaches to emotion*, 3:19–344, 1984.

[4] M. S. Goodwin, J. Groden, W. F. Velicer, L. P. Lipsitt, M. G. Baron, S. G. Hofmann, and G. Groden. Cardiovascular arousal in individuals with autism. *Focus on Autism and Other Developmental Disabilities*, 21(2):100–123, 2006.

[5] F. Grond, T. Bovermann, and T. Hermann. A supercollider class for vowel synthesis and its use for sonification. In *17th Annual Conference on Auditory Display (ICAD2011)*, 2011.

[6] T. Hermann. Taxonomy and definitions for sonification and auditory display. In *Proceedings of the 14th International Conference on Auditory Display (ICAD2008)*, 2008.

[7] T. Hermann, G. Baier, U. Stephan, and H. Ritter. Kernel regression mapping for vocal EEG sonification. In *Proceedings of the 14th International Conference on Auditory Display (ICAD2008)*, 2008.

[8] T. Hermann, K. Bunte, and H. Ritter. Relevance-based interactive optimization of sonification. In *Proceedings of the 13th International Conference on Auditory Display (ICAD2007)*, 2007.

[9] R. P. Hobson. The autistic child's appraisal of expression of emotion. *Journal of Child Psychology and Psychiatry*, 27(3):321–342, 1986.

[10] S. Koelsch, T. Fritz, D. Y. Cramon, K. Müller, and A. D. Friederici. Investigating emotion with music: An fmri study. *Human Brain Mapping*, 27:239–250, 2006.

[11] S. Namba, S. Kuwano, T. Hashimoto, B. Berglund, Z. D. Rui, A. Schick, H. Hoege, and M. Florentine. Verbal expression of emotion impression of sound. A cross-cultural study. *Journal of the Acoustical Society of Japan*, 12(1):19–29, 1991.

[12] R. W. Picard. Future affective technology for autism and emotion communication. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1535):3575–3584, 2009.

[13] J. Posner, J. A. Russell, and B. S. Peterson. The circumplex model on affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 2005.

[14] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[15] K. R. Scherer. Which emotions can be induced by music? what are the underlying mechanisms? and how can we measure them? *Journal of New Music Research*, 33(3):239–251, 2004.

[16] M. Schröder. Emotional speech synthesis: A review. *Interspeech*, pages 561–564, 2001.

[17] E. Schubert, S. Ferguson, N. Farrar, and G. E. McPherson. Sonification of emotion I: Film music. In *17th International Conference on Auditory Display (ICAD2011)*, 2011.

[18] M. Tachibana, J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi. Hmm-based speech synthesis with various speaking styles using model interpolation. In *Speech Prosody, International Conference*, 2004.

[19] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 1995.

[20] R. M. Winters and M. M. Wanderley. Sonification of emotion: Strategies for continuous display of arousal and valence. In *Proceedings of the 3rd International Conference on Music & Emotion (ICME3)*, 2013.