The Perceived Emotion of Isolated Synthetic Audio: The EmoSynth Dataset and Results

Alice Baird
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, Universität Augsburg, Germany.
alice.baird@informatik.uni-augsburg.de

Emilia Parada-Cabaleiro
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, Universität Augsburg, Germany.
emilia.parada-cabaleiro@informatik.uni-augsburg.de

Cameron Fraser
The Center for Digital Arts and Experimental Media, University of Washington, USA.
cpfraser@uw.edu

Simone Hantke
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, Universität Augsburg, Germany.
MISP Group, MMK Technische Universität München, Germany.
simone.hantke@informatik.uni-augsburg.de

Björn Schuller
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, Universität Augsburg, Germany
GLAM – Group on Language, Audio, and Music, Imperial College London, UK.
schuller@IEEE.org

ABSTRACT

The ability of sound to enhance human wellbeing has been known since ancient civilisations, and methods can be found today across domains of health and within a variety of cultures. There are an abundance of sound-based methods which show benefits for both physical and mental-states of wellbeing. Current methods vary from low frequency vibrations to high frequency distractions, and from drone-like sustain to rhythmical pulsing, with limited knowledge of a listeners psycho-physical perception of this. In this regard, for the presented study 40 listeners were asked to evaluate the perceived emotional dimensions of Valence and Arousal from a dataset of 144 isolated synthetic periodic waveforms. Results show that Arousal does correlate moderately to fundamental frequency, and that the sine waveform is perceived as significantly different to square and sawtooth waveforms when evaluating perceived Arousal. The general results suggest that isolated synthetic audio can be modelled as a means of evoking affective states of emotion.

AM'18, September 12–14, 2018, Wrexham, United Kingdom

ACM ISBN 978-1-4503-6609-0/18/09…$15.00

https://doi.org/10.1145/3243274.3243277

CCS CONCEPTS

• **Applied computing** → *Sound and music computing*; • **Human-centered computing** → Sound-based input / output;

KEYWORDS

affective computing, synthetic audio, sound healing, perception, wellbeing.

ACM Reference Format:
Alice Baird, Emilia Parada-Cabaleiro, Cameron Fraser, Simone Hantke, and Björn Schuller. 2018. The Perceived Emotion of Isolated Synthetic Audio: The EmoSynth Dataset and Results . In *Audio Mostly 2018: Sound in Immersion and Emotion (AM'18), September 12–14, 2018, Wrexham, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3243274.3243277

1 INTRODUCTION

The soundscape of our world is informative and vibrant, able to create cultural snapshots [21] and evoke emotional connections [32]. In the natural world sound is all encompassing, with parameters such as amplitude, and fundamental frequency continually changing. On one hand, sound (particularly music), has shown to have the ability to alter a human's state-of-conciousness [1], yet on the other hand, sound (particularly stochastic sources), can not only cause long-term physical health issues, such as hearing-loss [3], but also negatively impact mental-health [28].

Traditional techniques including *medicine songs* (i. e., chanting [11]) have been used by ancient cultures for healing, and these methods are still in practice within many communities across the globe, from Native North and South American tribes, to Korean Buddhist temples [11]. With the substantial 'power' of sound known to some degree, much focus is being put towards the development of systems to reduce the affect

of uncontrolled sound environments, e. g., road noise [31]. The implementation of purposeful soundscape designs receives less attention, and system based on holistic methods may have benefits to an abundance of scenarios.

In this way, synthetic audio generated through signal processing methods is gaining in clinical recognition [12], with sound-based vibratory apparatus (also known as *vibroacoustics* showing benefits for chronic pain e. g., arthritis [25]. Low frequency oscillations in particular, have been suggested to improve negative-mood, and reduce short-term stress [43], and sufferers of neurological disorders such as *Thalamocortical Dysrhythmia* are provided with low-frequency sound stimulation as a regulatory function [36]. Although specific use-cases in this area have been evaluated, previously no general assessment of fundamental synthetic audio features, including sound intensity, and overall duration have been made in relation to emotion.

The current state-of-the-art for computer audition frameworks (e. g., WaveNet [44]) shows great promise for soundscape design informed by additional attributes. With this in mind, this study asked 40 German native individuals to evaluate the perceived emotional dimensions of Valence and Arousal from a dataset 144 isolated audio instances, to gain a basis for future tasks in the area of audio generation. The Emotional Synthetic Audio (EmoSynth) dataset has been prepared specifically for this study, and includes audio of varied acoustic parameters – waveform, amplitude envelope, duration, and fundamental frequency. To this end, this study aims to gain a base understanding of the extent to which emotion is perceived in varied combinations of isolated synthetic audio.

2 RELATED RESEARCH

The Impact of the Soundscape

The notion of the soundscape is a term which was initially understood as an ecological concept in which culture and history can be captured, analogous to archival photography [42]. On the other side, the soundscape can include purposefully designed sound combinations, which may fill an environment for a specific intention e. g., for environments in which excessive background noise is causing task-distraction. Showing a substantial effects on wellbeing for individuals across cultures [33], the study of soundscapes is a cross-disciplinary topic, from environmental noise management [8] to cultural anthropology [9]. Cross culturally the emotional connections that different acoustic soundscapes evoke, predominately through association, have also recently been evaluated [32]. Moscoso *et al.* found that rural sounds often have a more positive impact than urban sounds, although in some cases, factors of modern-life such as rainforest deforestation will impact negatively. Additionally, synthetic soundscapes based on physiological signals (e. g., the heart beat), have been found to evoke much stronger emotions than purely synthetic audio [39].

Health and Wellbeing from Sound

Within the field of audio and acoustics there have in the past been many studies focusing on the effect of exposure to sound, both in the home and in the workplace [24]. Some studies claim that excessive sound level can have an effect on the hospital working environment, causing long-term implication for nursing staff [37]. In that same domain, music within a hospital has shown to have a strong impact on patient experience [46]. Health and wellbeing from music specifically, is a prominent area of research, with the strong links between them still not fully understood [30]. Vocal stimulation, through singing, is another method to maintain a positive level of both emotional and social wellbeing [38]. With non-verbal inward singing techniques, known as 'toning', and 'overtoning' able to resonate the brain and improve the flow of cranial fluid [12].

Synthetic audio and the design of synthetic soundscapes have been explored across domains of research, for their ability to promote wellbeing, and have shown positive results for the reduction of stress in public settings including the hospital emergency room [10]. Additionally, it has been shown that for self-reported anxiety, embedded binaural beats can clinically reduce such values [45].

Assessment of Emotion from Sound

The study of emotion crosses many domains of research, and many standardised measures have been developed for this type of assessment, particularly for the field of music [20]. There are 3 core approaches for this: *dimensional*, *continuous*, and *categorical*. The Self-Assessment Manikin (SAM) is one such dimensional approach. SAM is a visual assessment technique in which the Arousal, Valence, and dominance of a given stimuli is measured. SAM has been implemented in [7], for the analysis of emotion from natural sound sources. Continuous measures overtime are also popular for the study of emotion in music [19]. For example, as music changes overtime, this method ensures that dimensions can be measured against specific markers. These markers could be points within a piece of music, composed to insight stronger emotion. Through a time-continuous approach such markers could be analysed for their effectiveness. Another approach would be a categorical approach in which a listener has a forced choice between a set of given labels.

Synthetic Audio Data

Large quantities of synthetic audio can be generated reasonably quickly, yet there are limited datasets available, and

seemingly none which have been annotated in terms of emotion. There are many large datasets of music recordings, such as the Million Song Dataset [4], or the RWC Music Database [15], with the labels limited to artist name, genre, release data etc. Some do feature a mood label for music listening categorisation, and an overview of such datasets can be found in [2]. The Nsynth dataset gathered for use with the *NSynth*: Neural Audio Synthesis framework [1] by the Google Research lab [14], is the first-of-its-kind aimed at Deep Learning tasks and provides more than 1 000 instruments with more than 300 000 note combinations. The Nsynth framework itself is designed to promote creativity in audio expression, and is a WaveNet-based [44] autoencoder for synthesizing audio. Such frameworks show great promise for the generation of soundscapes, and with large quantities of data and training time, these show high fidelity in the domains of music and speech synthesis generation are possible [40]. In this way, the potential to generate emotion driven audio could be the next step, and the data set discussed within this study would be novel in that regard.

3 METHODOLOGY

The Emotional Synthetic Audio (EmoSynth)Dataset

As a means of evaluating the emotional dimensions evoked by isolated synthetic audio, a dataset of audio files has been generated using the computer programming language Csound [27]. For this initial study, the dataset includes 3 x waveforms, 12 x frequencies, 2 x amplitude envelop lengths, and 2 x durations. The data annotated data is publicly available[2], and includes 144 files (44.1 kHz, 16 bit WAV files), at a total length of 18 minutes. There have been previous studies which explore more complex manipulated waveforms [10]; however, this study focuses on a limited selection of acoustic parameters, as a means of providing a base for further, more complex study.

The **waveforms** sine, square, and sawtooth have been selected for the dataset, all of which are periodic waveforms. These have been chosen due to their wave shape (cf. Figure 1) and spectral variance.

The *sine* wave (also known as a sinusoidal wave, or pure tone), is a continuous, smooth periodic oscillation. The spectrum of a sine wave consists of a fundamental frequency (without harmonics), and has been described to lack 'timbre in the same sense that white lacks color' [6]. With such *pure* qualities, this waveform offers a baseline in which more complex combination can then be considered from.

[1] *Nsynth* Neural Audio Synthesis https://magenta.tensorflow.org/nsynth
[2] The Emosynth dataset is freely available for research purposes. Please contact the corresponding author to gain access.

Table 1: Frequency values considered within the dataset. Frequency Class (*fc*) and fundamnetal frequency (Hz), the Pitch Class (*pc*), and musical (N)ote.

fc	Hz	*pc*	N	*fc*	Hz	*pc*	N
1	41.20	1	E1	7	466.16	6	A#5
2	61.73	8	B1	8	698.45	3	F5
3	92.49	2	F#2	9	1046.50	10	C6
4	138.59	9	C#3	10	1567.98	5	G6
5	207.65	4	G#3	11	2349.32	12	D7
6	311.12	11	D#4	12	3520.00	7	A7

The *square and the sawtooth* waves are both non-sinusoidal periodic waveforms. The square has instantaneous transitions, and the sawtooth a sharp ramp-up and release. Spectrally, they are both created through the addition of sinusoidal waves [35], with the square wave showing a sum of only odd-integer harmonic frequencies, and the sawtooth showing all-integer harmonic frequencies [22]. Additionally the triangle wave, would include only even-integer harmonic frequencies. The triangle wave has been excluded for this analysis, as we focus on the impact of increased harmonic content on emotional perception, and the triangle waveform would in that sense be similar to square waveform.

There are 12 **fundamental frequency** values in the data set, which have been selected following the circle of fifths, at *Perfect 5th* intervals (7 semitones) from E1 – A7 (Table 1). This was chosen in order to equally consider a range of pitch classes (i. e., C–B) typical of western musical theory. According to the 'Docterin of Ethos' from Ancient Greece, specific scales of pitch class can evoke emotional and moral concepts [16]. Given that the fundamental frequencies used here do cover specific pitch classes, we might hypothesise that the perception of pitch class to some extent relates to emotional concepts. However, the consonant *Perfect 5th* intervals alone will not have an impact on the listener's perception of emotion, as audio instances were randomised during playback, and annotations are made directly after listening to only one audio instance. In this way, listeners are not able to evaluate the interval relationships – only the absolute value of the frequency.

The **amplitude envelope** (ae) parameters of attack and release were also applied to the signal. An amplitude envelope is used to control the time it takes for a waveform to reach its peak amplitude (attack), decaying to a steady-state of sustain, and then releasing to null [13], (commonly, an attack, decay, sustain (ADSR) envelope). For this study, attack and release are applied with two variable lengths 0.4*s* (ae-short) and 1.9*s* (ae-long), and all sound durations have both variable lengths applied. Amplitude Envelope has been

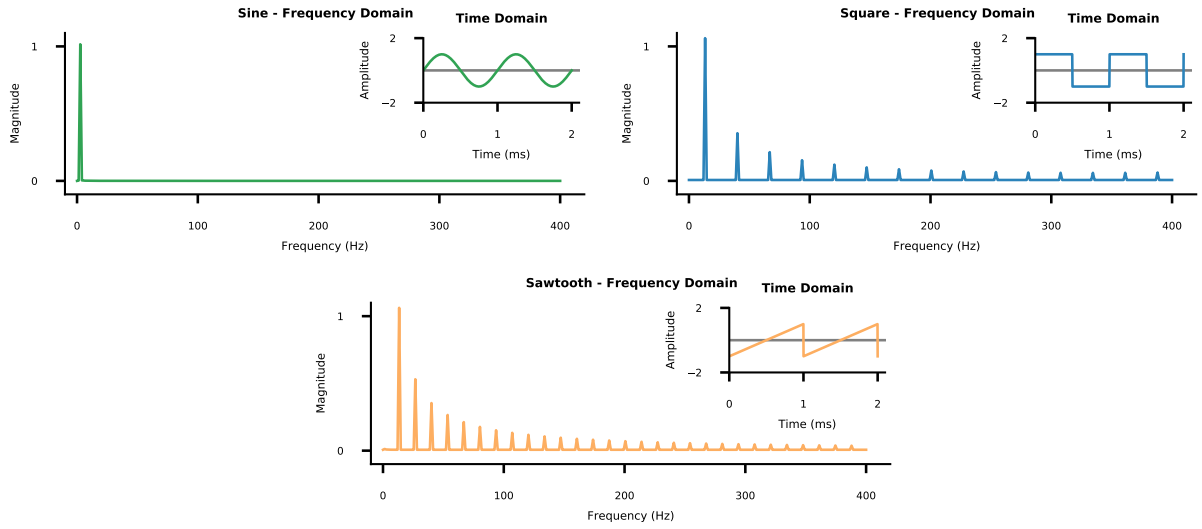

Figure 1: Time and frequency domain representations of the waveforms generated for the EmoSynth dataset. Sine, Square, and Sawtooth periodic waveforms.

discussed in the field of phonetics as having an explicit effect on a listener's ability to perceive emotion [29]; to this end, we evaluate the extent to which ae may influence the perception of emotion evoked by synthetic audio.

The **duration** of the audio files was also generated with two variations, 5 seconds (du-short) and 10 seconds (du-long). Although in the past, duration of music listening has shown to have little effect on the perception of emotion by listeners [5], there is seemingly no studies which have focused on isolated synthetic audio. Therefore, for the first time, we evaluate the relationship between perceived emotional states and isolated synthetic audio listening duration.

Perception Test Parameters and Set-up

For this study we are focusing on the perceived emotion that is evoked by the the instances of audio. Perceived emotion refers to the intellectual processing ability of the listener, and alternatively measuring induced emotion refers to the unconscious physiological response shown, which therefore would require additional measuring tools, such as skin conductance or brain activity [23].

As previously discussed in section 2 *Assessment of Emotion from Sound*, there are a few approaches which could have been used for the evaluation of emotion from synthetic audio. As categorical measures can cause ambiguous results [18], we evaluate with a dimensional model of emotion, considering *Arousal* and *Valence* (common in the field perception of emotion evaluation [41]). Valence is a measure of how positive or negative an emotion is, and Arousal measures the emotions levels of activation (i. e., weak, or strong intensity). For each dimension, i. e., Valence / Arousal, we consider a

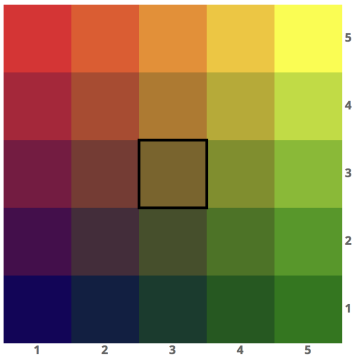

Figure 2: The interface used within iHEARu-PLAY, to label the audio for its perceived level of evoked Valence (x) and Arousal (y). Black outline box indicates current selection.

5-level scale, from 1 = negative / weak, 5 = positive / strong). We choose not to use a time-continuous dimensional model, as the instances do not change substantially over time, and due to possible 'annotator delay', these instances would be too short for an accurate evaluation of this.

A group of 40 (20 females and 20 males) German native listeners, aged between 21 – 29 years (mean, 23 years), voluntarily evaluated the dataset of 144 audio files (total duration 18 minutes). Listeners were asked to perform the task alone in a self-reported calm environment, in this way ensuring prior mood and or emotional state would have a limited influence on the results. The listening task was completed in the iHEARu-PLAY online browser-based annotation platform [17], and listeners were asked to label on a 5x5 grid the values of Valence (x-axis) and Arousal (y-axis), i. e., choosing

one unique position of the matrix as the emotional value of that sample (cf. Figure 2). Listeners were required to use headphones, and for safety purposes prior to the listening test, listeners set their volume carefully for the highest and lowest fundamental frequencies in the dataset, and were explicitly asked not to adjust their volume during the test. Before the main annotation began, listeners also made a practice run annotation, to ensure they were comfortable with the task, and understood the parameters being evaluated.

4 EVALUATION OF RESULTS

To evaluate the perceived Valence and Arousal, we analyse the results by their separate parameter (cf. Table 2). In the analysis, Cohen's d (for parametric test, i. e., *two-way ANOVA*), and η^2 (for non-parametric, i. e., *Kruskal-Wallis*) are considered as measures of effect size. Such values must be interpreted as follows [26]: $d = 0.2$ (small), $d = 0.5$ (medium), $d = 0.8$ (large); and $\eta^2 = 0.01$ (small), $\eta^2 = 0.06$ (medium), $\eta^2 = 0.14$ (large). The Pearson Correlation Coefficient (r) should be interpreted as [34]: .900 − 1.00 (very high), .700 − .900 (high), .500 − .700 (moderate), .300 − .500 (low), .000 − .300 (negligible). Finally, here, p values indicate statistical significance above the conventional threshold of $p < .05$. However, since the sample size might influence these parameters, analysis will mainly focus on the effect size and correlation results.

Dimensional perception of amplitude envelope and duration

As the parameters of amplitude envelope and duration are both time dependent (i. e., short, long), we evaluate them together. The sine waveform has been chosen for this evaluation, as this wave is a pure tone which we see as our baseline. For both the perception of Arousal and Valence in the isolated synthetic audio, a two-way ANOVA test was considered in order to examine the effect of duration (du), i. e., du-short and du-long, and amplitude envelope (ae), i. e., ae-short and ae-long.

(i) Arousal assessment – A two-way ANOVA has been performed for each listener group (female and male), showing in both cases that there is no statistically significant interaction between the two independent variables (du and ae) and the dependent variable (Arousal perception). For female and male listeners, the test yielded to F(1,960) = 3.718, $p = .054$; and F(1,960) = 0.335, $p = .563$ respectively. Thus, perception of Arousal was evaluated with all the listeners together, showing no significant results from the two-way ANOVA: F(1,1920) = 0.018, $p = .892$.

Simple main effect analysis has shown that *ae* does not influence listeners' perception of Arousal. For the evaluation of *du-long* and *du-short* samples, there is also no significance ($p = .122$, effect size $d = 0.10$), and ($p = .191$, effect size $d = 0.09$) respectively. *du* is also similarly perceived regardless of *ae*, as shown by the lack of significance for both *ae-long* ($p = .093$, effect size $d = 0.12$), and *au-short* ($p = .057$, effect size $d = 0.13$).

Table 2: The mean and standard deviation (sd) result for valence and Arousal from each parameter in the dataset. Waveforms (Wave), Frequency Class (*fc*), Amplitude Envelope (*ae*), and Duration (*du*).

Wave	**Valence**	*sd*	**Arousal**	*sd*
Sine	2.97	1.32	2.85	1.05
Square	2.93	1.34	3.55	1.24
Sawtooth	2.84	1.31	3.58	1.09
fc	**Valence**	*sd*	**Arousal**	*sd*
1	2.20	1.08	2.28	1.02
2	2.14	1.01	2.50	1.09
3	2.47	1.15	2.75	1.14
4	2.81	1.16	2.93	1.06
5	3.19	1.26	3.13	1.08
6	3.77	1.19	3.43	1.16
7	2.19	1.12	3.28	0.98
8	2.43	1.17	3.45	0.99
9	2.84	1.24	3.83	0.99
10	3.24	1.25	3.76	1.01
11	3.58	1.26	4.08	0.85
12	3.82	1.26	4.30	0.86
ae	**Valence**	*sd*	**Arousal**	*sd*
ae-short	2.87	1.34	3.44	1.17
ae-long	2.95	1.30	3.23	1.17
du	**Valence**	*sd*	**Arousal**	*sd*
du-short	2.90	1.31	3.24	1.17
du-long	2.92	1.34	3.43	1.18

(ii) Valence assessment – A two-way ANOVA was also conducted in order to examine the effect of *du* and *ae* in the listeners' perception or Valence. For both female and male our analysis shows that there is a statistically significant interaction between the influence of these two parameters and the assessment of Valence. For female, the two-way ANOVA yielded to F(1,960) = 12.229, $p = .001$; for male to F(1,960) = 4.694, $p = .030$. Therefore, since both female and male listeners show an influence in Valence for duration and amplitude envelope, all the responses were again evaluated together, yielding to F(1,1920) = 7.195, $p = .007$.

Simple main effect analysis has shown that *du* does not play a role in listeners' perception with *ae-long* ($p = .075$, effect size $d = 0.12$); whereas, it is relevant for the evaluation of the *ae-short* samples ($p = .045$, effect size $d = 0.13$). *ae* seems to affect listeners' perception of Valence for *du-long* samples ($p = .008$, effect size $d = 0.18$); but this has not been

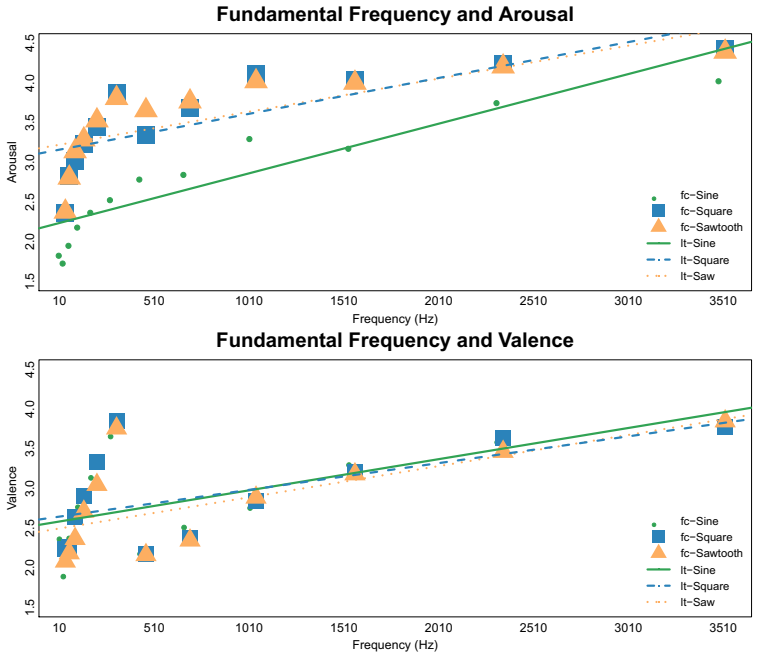

Figure 3: Mean results for Arousal and Valence of each Frequency Class (fc), by each waveform type (Sine, Square and Sawtooth). The linear trend (lt) line is also shown for each waveform.

confirmed for the evaluation of *du-short* samples ($p = .267$, effect size $d = 0.07$). The small effect size shown by the single main effects suggests that neither *du* nor *ae* are factors of influence for the perception of Valence.

Dimensional perception of the waveform

For the evaluation of waveforms (sine, sawtooth, and square) in relation to the emotion evoked, the annotations for each waveform have been comparatively evaluated, considering the two emotional dimensions, i. e., Arousal and Valence, individually. Considering that *du* and *ae* did not show consistently to have an effect on the dimensional perception, all the samples have been evaluated regardless of *du* or *ae*. The non-parametric test *Kruskal-Wallis* has been performed for the evaluation of both Arousal and Valence since the null hypothesis for the data homogeneity (measured by *Levene* test) and normality (measured by *Kolmogorov-Smirnov* test) has been rejected.

For Arousal, the variance of the populations from the different groups (i. e., the three wave forms), were not equal or normal; yielding for Levene to $F(2,5760) = 29.011$, $p < .001$ and for Kolmogorov-Smirnov to $p < .001$. For Valence, even though the null hypothesis for homogeneity was confirmed, yielding in Levene test to $F(2,5760) = 1.223$, $p = .294$ and for Kolmogorov-Smirnov to $p < .001$, thus rejecting the null hypothesis of normality.

(i) Arousal assessment – Considering this, the non-parametric test *Kruskal-Wallis* has been performed, which has shown that the Arousal evoked by the three waveforms is perceived as different to a statistically significant level: $H(2) = 404.101$, $p < .001$. In order to evaluate of which specific waves differences were displayed, a pairwise comparison, considering the post hoc test Dunn-Bonferroni was applied.

The pairwise comparison shows that the sine wave is perceived significantly different to the sawtooth and square wave. This comparison displayed in both a p value $< .001$ and a medium effect size $\eta^2 = 0.08$. On the contrary, the comparison of the sawtooth wave vs square wave shows that these two waves are not perceived as significantly different ($p = 1.00$, and an almost none effect $\eta^2 < 0.001$).

(ii) Valence assessment – From the non-parametric *Kruskal-Wallis* test, the results have shown that the three wave forms are perceived differently in terms of Valence to a statistically significant level: $H(2) = 9.381$, $p = .009$. However, even though the pairwise comparison shows that sine wave is perceived statistically different than the sawtooth wave (p value $= .012$), the very small effect size $\eta^2 = 0.002$ indicates that such difference might be negligible. Furthermore, no significant difference in the perception has been shown in the other pairwise comparisons, i. e., sine vs sawtooth ($p = 1.00$, $\eta^2 < 0.001$ for sine vs square, and $p = .062$, $\eta^2 = 0.001$ for sawtooth vs square).

Dimensional perception of pitch

Linear relationship between fundamental frequency and perceived Arousal and Valence has been evaluated for the three waveforms (sine, sawtooth, and square) independently, as we have seen some significance to their effect. For Arousal, a 2-tailed Pearson correlation test yielded a moderate positive correlation ($r = .516$ and $p < .001$) for sine waves, i. e., the higher the fundamental frequency, the higher the perceived Arousal (cf. Figure 3). On the contrary, low correlation has been displayed for the other two waveforms: $r = .411$ ($p < .001$) for sawtooth and $r = .434$ ($p < .001$) for square wave. For Valence, the correlation with fundamental frequency is low, yielding to $r = .300$ ($p < .001$) for sine wave, to $r = .311$ ($p < .001$) for sawtooth, and to a negligible correlation of $r = .270$ ($p < .001$) for square wave.

From the lower plot of Figure 3, it can be observed that the fundamental frequency against Valence may in-fact relate to melodic pairings of the pitch classes rather than frequency class. In order to assess this correlation for both Valence and Arousal, the fundamental frequencies have been reordered according to the chromatic scale, (cf. Table 1 column *pc*), and a 2-tailed Pearson correlation test has been performed. Yet, from this we see that neither Arousal or Valence correlate against pitch classes, yielding negligible correlations: sine waves, $r = .211$ ($p < .001$) for Arousal and $r = .189$ ($p < .001$) for Valence; sawtooth waves, $r = .271$ ($p < .001$) for Arousal and $r = .260$ ($p < .001$) for Valence; square waves, $r = .307$ ($p < .001$) for Arousal and $r = .241$ ($p < .001$) for Valence.

5 CONCLUSIONS

From this study an understanding of the dimensional emotional measure of Valence and Arousal percieved in a dataset of isolated synthetic audio with varied acoustic parameters was gained. The perception of Valence does not show any significant results, however Arousal shows more significance with fundamental frequency correlating positively. The sine wave shows to be significantly different to all other waveforms in the dataset particularly for Arousal. This suggests that for the creation of a low aroused soundscapes, lower fundamental frequencies with less complex wave combinations may be more suitable. However, as [39] found that the presence of physiological signals within a soundscape can enhance emotional perception, we will consider exploring this combination as a future step, as well as evaluating the trends found here further. Additionally, with synthetic audio showing the possibility to evoke emotional dimensions, further study will include developing longer audio combinations and incorporating aspects such as spatial panning. As well as this, from the generation of a much larger dataset, transfer learning methods can be used in order to expand the annotation labels obtained in this study, making deep learning methods a more feasible option for audio generation.

ACKNOWLEDGMENTS

This work is supported by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B), and the European Union's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu).

REFERENCES

[1] Ludwig A.M. 1966. Altered states of consciousness. *Archives of General Psychiatry* 15, 3 (1966), 225–234.
[2] Mathieu Barthet, György Fazekas, and Mark Sandler. 2013. Music Emotion Recognition: From Content- to Context-Based Models. In *From Sounds to Music and Emotions*, Mitsuko Aramaki, Mathieu Barthet, Richard Kronland-Martinet, and Sølvi Ystad (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 228–252.
[3] Mathias Basner, Wolfgang Babisch, Adrian Davis, Mark Brink, Charlotte Clark, Sabine Janssen, and Stephen Stansfeld. 2014. Auditory and non-auditory effects of noise on health. *The Lancet* 383, 9925 (2014), 1325 – 1332.
[4] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. (2011), 16 pages.
[5] E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet. 2005. Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition and Emotion* 19, 8 (2005), 1113–1139.
[6] Meera M. Blattner, Denise A. Sumikawa, and Robert M. Greenberg. 1989. Earcons and Icons: Their Structure and Common Design Principles. *Human Computer Interaction* 4, 1 (1989), 11–44.
[7] Margaret M Bradley and Peter J Lang. 2000. Affective reactions to acoustic stimuli. *Psychophysiology* 37, 2 (2000), 204–215.
[8] A. L. Brown. 2010. Soundscapes and environmental noise management. *Noise Control Engineering Journal* 58, 5 (2010), 493–500.
[9] A. L. Brown, Jian Kang, and Truls Gjestland. 2011. Towards standardization in soundscape preference assessment. *Applied Acoustic* 72, 1 (2011), 387–392.
[10] David Brown. 2012. *Designing sound for health and wellbeing.* RMIT University, Melbourne, Australia. 1–120 pages.
[11] P M. Cook. 1995. Sacred wellness: Music and healing among indigenous people.. In *Proc. of Annual Conference of the National Association for Music Therapy Western Region.* Colorado State University, Seattle, WA, USA, 891–897.
[12] Barbara J. Crowe and Mary Scovel. 1996. An Overview of Sound Healing Practices: Implications for the Profession of Music Therapy. *Music Therapy Perspectives* 14, 1 (1996), 21–29.
[13] Charles Dodge and Thomas A. Jerse. 1997. *Computer Music: Synthesis, Composition and Performance* (2nd ed.). Macmillan Library Reference, New York City, New York, USA.
[14] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. 2017. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. (2017), 16 pages.
[15] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. 2011. RWC Music Database: Music Genre Database and Musical Instrument Sound Database. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR).* ISMIR, Miami, Florida, USA, 229–230.
[16] Donald J. Grout and Claude V. Palisca. 2001. *A history of Western music.* Norton, New York, NY, USA.
[17] Simone Hantke, Florian Eyben, Tobias Appel, and Björn Schuller. 2015. iHEARu-PLAY: Introducing a game for crowdsourced data collection

for affective computing. In *Proc. 1st International Workshop on Automatic Sentiment Analysis in the Wild (WASA 2015) held in conjunction with the 6th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2015)*. AAAC, IEEE, Xi'an, P. R. China, 891–897.

[18] Hu, X. and Downie, J. S. and Laurier, C. and Bay, M., and Ehmann, A. F. 2008. The 2007 MIREX audio mood classification task: Lessons learned. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*. ISMIR, Philadelphia, PA, USA, 462–467.

[19] P. N. Juslin and P. Laukka. 2004. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research* 33 (2004), 217–238.

[20] Patrik N. Juslin and Daniel Västfjäll. [n. d.]. Emotional Responses to Music: The need to consider underlying mechanisms. *Journal of Behavioural and Brain Sciences* 31 ([n. d.]), 559–621.

[21] Jian Kang and Brigitte Schulte-Fortkamp. [n. d.]. *Soundscape and the Built Environment.* Florida, USA.

[22] Ulrich Karrenberg. 2007. *An interactive multimedia introduction to signal processing*. Springer Science & Business Media.

[23] S. Khalfa, P. Isabelle, B. Jean-Pierre, and R. Manon. 2002. Event-related skin conductance responses to musical emotions in humans. *Neuroscience Letters* 328, 2 (2002), 145–149.

[24] V.J. Krichagin. 1978. Health effects of noise exposure. *Journal of Sound and Vibration* 59, 1 (1978), 65 – 71.

[25] R. M. Chirieac L. M. Ailioaie, C. Ailioaie and A. Chirieacy. 2011. Effects of physical and vibroacoustic therapy in chronic pain in juvenile arthritis. *Revista Romanade Reumatologie* 20, 3 (2011), 198âĂŞ–202.

[26] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology* 4 (2013), 863.

[27] Victor Lazzarini, Steven Yi, John Fitch, Joachim Heintz, Oyvind Brandtsegg, and Iain McCurdy. 2016. *Csound: A Sound and Music Computing System* (1st ed.). Springer Publishing Company, Incorporated.

[28] P Lercher, G W Evans, M Meis, and W W Kofler. 2002. Ambient neighbourhood noise and children's mental health. *Occupational and Environmental Medicine* 59, 6 (2002), 380–386.

[29] Philip Lieberman and Sheldon B. Michaels. 1962. Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech. *The Journal of the Acoustical Society of America* 34, 7 (1962), 922–927.

[30] Raymond A. R. MacDonald, Gunter Kreutz, and Laura Mitchell. 2013. *Music, Health, and Well-being.* Oxford University Press. 1–550 pages.

[31] Ghorbanali Mohammadi. 2009. An Investigation of Community Response to Urban Traffic Noise. In *Global Perspective for Competitive Enterprise, Economy and Ecology*, Shuo-Yan Chou, Amy Trappey, Jerzy Pokojski, and Shana Smith (Eds.). Springer London, London, 673–680.

[32] Paola Moscoso, Mika Peck, and Alice Eldridge. 2011. Emotional associations with soundscape reflect human-environment relationships. *Journal of Ecoacoustics* 2, 1 (2011), 1–19.

[33] Paola Moscoso, Mika Peck, and Alice Eldridge. 2018. Systematic literature review on the association between soundscape and ecological/human wellbeing. (2018), 28 pages.

[34] Mavuto M Mukaka. 2012. A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal* 24, 3 (2012), 69–71.

[35] Meinard Müller. 2015. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications* (1st ed.). Springer Publishing Company, Incorporated.

[36] Lili Naghdi, Heidi Ahonen, Pasqualino Macario, and Lee Bartel. 2015. The effect of low-frequency sound stimulation on patients with fibromyalgia: A clinical study. *Pain research and management* 20, 1 (2015), 21–27.

[37] Selen Okcu, Erica E. Ryherd, Craig Zimring, and Owen Samuels. 2011. Soundscape evaluations in two critical healthcare settings with different designs. *The Journal of the Acoustical Society of America* 130, 3 (2011), 387–392.

[38] Antoinette Olivier and Hetta Potgieter. 2015. Create music that will open a person's heart": a perspective on emotional and social wellbeing as depicted in three films. *Koersjournal* 80, 4 (2015), 1–8.

[39] Emilia Parada-Cabaleiro, Alice E. Baird, Nicholas Cummins, and Björn Schuller. 2017. Stimulation of Psychological Listener Experiences by Semi-Automatically Composed Electroacoustic Environments. In *Proceedings 18th IEEE International Conference on Multimedia and Expo, ICME 2017*. IEEE, IEEE, Hong Kong, P. R. China, 1051–1056. (acceptance rate: 30 %, IF* 0.88 (2010)).

[40] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2018. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. (2018), 16 pages.

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

[42] R Murray Schafer. 1993. *The soundscape: Our sonic environment and the tuning of the world.* Vermont, USA.

[43] Olav Skille. 1989. VibroAcoustic Therapy. *Music Therapy* 8, 1 (1989), 61–77.

[44] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. *CoRR* abs/1609.03499 (2016), 4.

[45] Tracey J Weiland, George A. Jelinek, Keely E. Macarow, Philip Samartzis, David M. Brown, Elizabeth M. Grierson, and Craig Winter. 2011. Original sound compositions reduce anxiety in emergency department patients: a randomised controlled trial. *The Medical Journal of Australia* 11, 5 (2011), 694–698.

[46] A. Zheng, R. Sakari, S. Cheng, A. Hietikko, P. Moilanen, J. Timonen, K. Fagerlund, M. Kaerkkaeinen, and M. Alen. 2009. Effects of a low-frequency sound wave therapy program on functional capacity, blood circulation, and bone metabolism in frail old men and women. (2009), 32 pages.