Interaction with the Soundscape: Exploring Emotional Audio Generation for Improved Individual Wellbeing

Alice Baird[1(✉)], Meishu Song[1], and Björn Schuller[1,2]

[1] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany
alice.baird@informatik.uni-augsburg.de

[2] GLAM – Group on Language, Audio, and Music, Imperial College London, London, UK

Abstract. Our daily interaction with the soundscape is in flux, and complex natural sound combinations have shown to have adverse implications on user experience. A computational approach to stabilise the sonic environment, tailored to a user's current affective state may prove beneficial in a variety of scenarios, including workplace efficiency, and exercise. Herein, we present initial perception test results, from a rudimentary approach for soundscape augmentation utilising chromatic feature sonification. Results show that arousal and valance dimensions of emotion can be altered through augmentation of three classes of natural soundscape, namely 'mechanical', 'nature', and 'human'. Proceeding this we outline a possible approach for an affective audio-based recognition and generation system, in which users (either individually or as a group within a specific environment) are provided with an *augmentation* of their current soundscape, as a means of improving wellbeing.

Keywords: Audio generation · Wellbeing · Machine learning · Human Computer Interaction

1 Introduction

The soundscape is the combined audio components being heard at a given moment in time [52]. Involuntarily, we are continually interacting with the soundscape, as – unlike visual interaction – we cannot 'close our ears' to stop an audible input. With this in mind, uncontrolled audio environments, have shown to impact individual wellbeing, substantially heightening stress, and causing a long-term decline in workplace efficiency [31].

In regards to these topics, there are many more research efforts occurring in the fields of affective acoustic ecology [15], and general sound recognition [12],

This work is funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

H. Degen and L. Reinerman-Jones (Eds.): HCII 2020, LNCS 12217, pp. 229–242, 2020.
https://doi.org/10.1007/978-3-030-50334-5_15

with smart-device applications for aiding sleep, and meditative states now being much more common[1]. However, such apps do not yet personalise the audio in real-time manner.

Herein, we present background and initial findings to support the development of a system which does create a personalised interaction with sound. Showing through previous studies that multiple modalities can be used non-intrusively to gain an understanding of a user's current state [3], and that deep generative approaches show the ability to generate affective data [4].

This contribution is structured as follows; first we conduct a brief literature review of related work under the topics of sound and stress reduction, Human Computer Interaction (HCI) and affective audio, and computational audio generation approaches. We then perform a preliminary perception study, based on a rudimentary approach for soundscape augmentation, and discuss the results. Proceeding this we propose a state-of-the-art method for the application of affective soundscape audio augmentation. Finally, we conclude our results, and summaries our outlook for the next steps of this research area.

2 Related Work

2.1 Sound for Wellbeing

When discussing sound as a taxonomy, this extends across many branches, from environmental sounds to speech. Within the field of *sound healing*, there are many sound sources including; acoustic and synthetic, which show to have a variety of wellbeing benefits including stress reduction [50]. Previous studies have suggested that excessive sound levels can have an effect on the hospital working environment, having long-term implications for nursing staff [40].

Acoustic-based tools are used commonly by healing practitioners, e. g. tuning forks at 128 Hz for relieving tissue-based abnormalities [19], or ritual communal drumming, which has shown to improve wellbeing in young people [59]. Another sound-based practice aimed at the reduction of stress is Transcendental Meditation (TM) [27], partially utilising the spoken mantra[2]. Additionally, through the integration of both sound and breathing techniques, TM has shown to both physiologically and psychologically reduce stress, quantitatively showing a decreased average theta (θ) when monitoring via Electroencephalography (EEG) and increased alpha (α) [18].

As well as such vocal mantras, practitioners of TM integrate a series of acoustic instruments such as the Tibetan or crystal singing bowls. These bowls have a long history of use in mediation [24] and are played with a continuous oscillation around the circumference of the bowl, resulting in a full overtone sound which

[1] Popular applications currently available for the purpose of aiding sleep and reducing states of arousal include: Headspace, Noisli, Pzizz, Slumber, Calm, Sleep Cycle, etc.

[2] The spoken mantra, through the repetition of phrases such as 'has no meaning', would be personal to the individual and is selected due to the resonant and harmonising ability within the meditator.

sustains a prolonged resonation. The Tibetan singing bowl has been applied to many stress reducing scenarios – including as an aid to school teachers [9] – and has shown to increase feelings of spirituality, in turn relieving symptoms of stress including tension [20]. The Tibetan Singing bowl has also been integrated in a variety of mHealth stress reducing targeted apps [22].

In regards to synthetic-based sound tools, there have been a variety of studies which have shown stress reduction results, in various environments [31]. Synthetic music within a hospital has shown to have a strong impact on a patient's experience [61]. Similarly, the acoustic environment of a workplace benefits from artificial acoustic design [30], and through integration of synthetically designed audio environments the workplace experience also improves [25]. Synthetic audio generation has also been investigated in the realm of therapeutic applications, specifically exploring how synthetic sound might influence listeners experiences in psychological areas, such as creativity or self-perception [44].

There is also much research focused on how listeners perceive music and how emotions are brought on by music or what psychological mechanisms causes these emotions [26]. For example, music is often used to enhance the emotional impact of movies [7]. Unlike most other stimuli that evoke emotions, such as encounters with dangerous animals, threats or facial expressions, music has no obvious, intrinsic survival value [34]. Blood et al. presented a novel approach to the study of music and emotion, using positron emission tomography to measure cerebral correlates of affective and perceptual responses to musical dissonance [8].

2.2 HCI and the Use of Audio for Wellbeing Applications

Over the past two decades, researchers have increasingly realised the importance of recognising the emotional aspects which occur during human-computer interaction (HCI) [11]. For example, in many HCI scenarios a computer aided tutoring system is highly desirable and a response based on emotional or cognitive state of the human user may improve user experience [55]. During interaction humans provide emotion-based cues from physical gesture, facial expressions and also the voice [23]. Nowadays affective recognition systems are mainly developed through 2 key qualitative steps: understanding emotional response, adapting the development based on user experience.

One method for non-intrusively understanding a user's experience is through the voice, and there is an abundance of HCI applications specifically in the realm speech recognition, e. g. voice dialling [43]. Automated speech recognition systems are also integrated in language learning paradigms to improve pronunciation [51]. As well as his Voice-based user interfaces are becoming ubiquitously available, being embedded both into everyday mobility via smartphones, and into the life of the home via assistant devices [46].

As well as the voice, there has been an increased interest in the impact of in-game audio. Paterson et al. developed an audio design with a complex and immersive soundscape, which is emotionally engaging and supports the game narrative [45]. Similarly, Roden et al. proposed a framework for interactive narrative-based audio only adventure games [47], and Sliwinski et al. explored

the development of an audio-visual game to induce wellbeing and mindfulness [56] Similarly, Rogers et al. discussed games which are considered relaxing and encourages research directions for exploring the role of game audio specifically, to improve player wellbeing, via stress reduction [48].

Thus, there is much research exploring the potential use of audio for wellbeing. Roger et al. identified the effect of music in games as a preventative measure against stress in everyday life by facilitation of relaxation [48]. In relation to health specifically, Willianmson et al. explored first-time mothers' breast-feeding difficulties through the use of audio-diaries [58], and Mirelman used audio-biofeedback for improving Parkinson's patients balance [38]. Additionally, Dijk et al. presented the concept of auditory–tactile stimulation for health and well-being through carefully selected audio–tactile stimuli causing a person's bodily, mental and emotional state to be altered [13].

2.3 Audio Generation

Although the scope of this study is focused largely on the generation of complex soundscapes, audio generation can refer to an array of audio-based fields, from speech synthesis to instrument modelling. In this regard, many of the methods mentioned will be found across all such domains, but are not limited to them. Conventional computational methods for audio generation would include a variety of digital signal processing approaches, such as Hidden Markov Models [53] or cellular automata [10]. These methods are still applied today, however the current state-of-the-art for the term audio generation would refer to a division within machine learning in which systems are largely data-driven [57].

An earlier deep approach for generating audio was Deep Minds WaveNet [57]. WaveNet is a progressive auto regressive generator, and is an audio adaptation of the PixelCNN [42], modelling features of raw audio which are represented as 8-bit audio files, with 256 possible values. During the training process, the model predicts values for waveforms (audio signals with a temporal resolution of at least 16 kHz samples per second) at each step comparing them to the true value, using cross-entropy as a loss function. In this way, the WaveNet architecture is applying a multi-class classification of 256-classes [35]. As a means of decreasing the computational time expense, that may be associated to such a classification task, WaveNet applies the method of stacked dilated casual convolutions, reducing the receptive field without any substantial loss in the resolution [60].

Although WaveNet has been showcased in the speech synthesis domain, the applications are broad. The original architecture showed promise for high fidelity in music with comparable human perception results [57]. Recently, an adaptation of the WaveNet framework is the NSynth (Neural Synthesizer) auto encoder specifically tailored towards synthesis of musical notes [16].

Another neural network approach, which was motivated by WaveNet , is SampleRNN [37]. This model is an unconditional end-to-end neural audio generation architecture that uses auto-regressive multilayer perceptron's and a Recurrent Neural Network (RNN), in a hierarchical structure, to capture temporal variance over large audio signal durations. Despite showing competitive human perception

results against WaveNet , the SampleRNN suffers from unrealistic computation time and the perception results are not shown to be significant, rather tendencies [37]. However, the advantages of time dependent RNNs would be suitable for soundscape generation offline.

First proposed by Goodfellow et al. in 2014 [21], Generative Adversarial Networks (GANs) have found recent popularity within the data generation domain and are arguably becoming a fundamental approach for this type of task. Essentially, generating new samples of audio based on raw audio signals, GANs are a pair of unsupervised networks which compete against each other, *generating* new instances of data until the *discriminator*, can no longer reliably tell a difference.

As well as being applied for the task of unsupervised representation learning from audio spectrograms [1], GANs aimed specifically for use with audio generation were first introduced in 2018, with WaveGANs and SpecGANS [14]. Approaches typically applied in the vision domain, were explored by extracting spectrogram images and comparing the networks ability to generate audible spectrogram instances. This was followed by the Conditional WaveGAN [33], which specifically focused on waveform generation through a concatenation based conditioning approach. Despite WaveGAN showing strong results for what is described as *human audible* samples, post-processing for noise reduction and appropriate optimisation due to instability were required.

3 First Step Soundscape Augmentation Perception Study

To evaluate the efficacy of augmentation of the original soundscape to alter emotional perception, we conducted a short listening test with 10 individuals[3]. Listeners evaluated arousal and valence dimensions of emotion [49], for each audio file (listening in a randomised order, twice before giving their score), on a 5-point Likert scale (e. g. 0 = Low arousal/valence, 4 = High arousal/valence). All listeners used headphones for this study.

3.1 Preliminary Acoustic Analysis

As we have mentioned previously in Sect. 2.1, the singing bowl is a common acoustic instrument used by healing practitioner (including in Transcendental Meditation) for improving states of wellbeing. With this in mind, we have chosen to use its most similar synthetic signal – a Sine wave – for this first-step augmentation approach. A sine wave (also known as a sinusoid) is a continuous periodic oscillation. As a function of time (t), a sine wave can be expressed as:

$$y(t) = Asin(2\pi ft + \varphi)$$

where in this case A refers to amplitude from zero, f the frequency, i. e. the number of oscillations (cycles) occur over t, and φ is *phase* i. e. when the cycle of oscillation is $t = 0$.

[3] 5 Female and 5 Male. Nationalities: 2 British, 4 Chinese, 4 German.

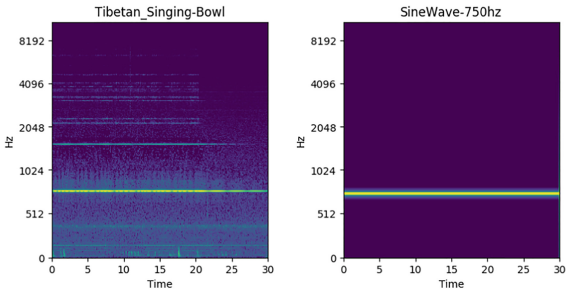

Fig. 1. Spectrogram representation of 30 s from a Tibetan Singing bowl recording (left), and a generated pure sine tone (right). Although similar in fundamental frequency, it should be noted, that overtones (as well as reverberation) which can be observed in the spectrogram representation of the singing bowl, may play a strong part in altering a listener's affective state.

We performed an initial acoustic analysis of multiple recordings from the singing bowl, taken from the Acoustics Sounds for Wellbeing Dataset [6], and compared this sine waves of matching frequency. Findings show that characteristics of the audio are similar (cf. Fig. 1, for spectrogram representation). For example, both are a continuous single frequency oscillation, and when monitoring pitch continuously the standard deviation came to 24.9 Hz, and 23.4 Hz for Tibetan and Sine, respectively. However, it is worth noting that aspects from the singing bowl such as resonance (and even human intervention) may play a deeper part in the improvement of wellbeing, and this is not replicated intrinsically through a single sine wave generator.

3.2 Audio Generation Approach

To summaries the rudimentary audio generation approach applied for this initial study, we utilised the Emo Soundscapes Database [17], and extracted Chroma features from 56 audio files (28 with lower rating of arousal and valence, and 28 with higher ratings of arousal and valence). Audio files were within the classes of 'Mechanical', 'Human', and 'Nature', and we then, sonified the corresponding chromatic notes (A-G#) as sine waves, overlaying this onto the original soundscapes.

To achieve this, we developed the first iteration of wellSounds[4] In this 'chromatic approach' we extract a 12 dimensional chromatic feature set from each trimmed (7 sec) audio file (prior to normalisation). At a given time-step based on the duration of the audio file. Features are then assigned to the corresponding Sine wave frequency (e. g. 65.4 Hz = C2, and 110.0 Hz = A2), and combined to make polyphonic (naive) chord combinations. The segments of audio are then concatenated to make a continuous 'augmentation' of the original audio file. The

[4] To apply the methods used in this study to new audio of fixed length visit the wellSounds Github: https://github.com/wellSounds/chromatic-approach.

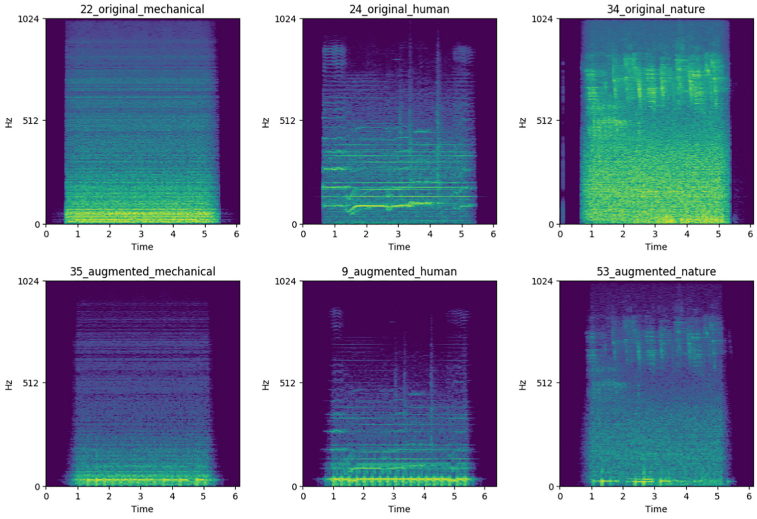

Fig. 2. Spectrogram representation of original audio and augmented audio. For each of the 3 classes – Mechanical, Human, and Nature. Through post processing of the original soundscapes it can be seen that the energy of noise is also reduced in the augmented soundscape, particularly prominent in the nature example.

resulting synthetic sine wave augmentation is then mixed onto the original audio file (proceeding a number of post-processing steps including equalisation and compression). A spectrogram representation of the WellSounds augmentation can be seen in Fig. 2[5].

3.3 Perception Study Results

Results from the study (based on the 3 classes), are shown in Table 1. To evaluate the significant (or not) difference between soundscape augmentation and original soundscape, we conduct a two-tailed T-test, rejecting the null-hypothesis at a significance level of $p < 0.05$ and below.

When observing the results from a class basis (cf. left of Table 1), of note we see there is a change in emotion perception across all classes, and particularly for the 'Nature' class a significant difference is shown between the augmented and original data types ($p = 0.001$, and 0.04 for valence and arousal, respectively). Although not necessarily a positive affect for the augmented soundscapes, this does show promise for the ability of such an augmentation approach to alter states of wellbeing. Additionally, from Fig. 3, we see that the standard deviation between listeners is quite wide, and therefore further studies with a larger group of listeners may give a more reliable trend.

[5] A selection of original and augmented soundscapes can be heard at the following link http://bit.ly/2T7uu4P.

Table 1. Results from perception study. Evaluating the perception of 10 listeners on a Likert scale of 0–4 for (V)alence and (A)rousal of the (Ori)ginal and (Aug)mented version of the soundscape. In the left table, results are presented based on the 3 soundscape classes (Mechanical, Nature, Human). In the right table, results are grouped by original ratings of (high) a (low) emotional dimensions of valence and arousal from the EmoSoundscape DB. Reporting Mean (μ) and Standard Deviation ($\pm$) across all listeners. * indicates significant difference, between (Ori)ginal soundscape, and (Aug)mented.

	Mechanical		Nature		Human	
	V	A	V	A	V	A
Ori (μ)	1.74	1.92	2.44	2.32	2.16	2.08
Aug (μ)	1.48*	2.05	1.88*	1.98*	1.55*	2.22
Ori ($\pm$)	0.93	1.11	1.17	1.09	0.92	1.03
Aug ($\pm$)	0.82	1.03	1.04	1.10	0.95	1.02

	High		Low	
	V	A	V	A
Ori (μ)	2.44	2.51	1.80	1.71
Aug (μ)	1.74	2.34	1.51	1.85
Ori ($\pm$)	1.18	1.04	0.81	0.99
Aug ($\pm$)	1.04	1.06	0.85	0.99

When looking at Table 1 (right) – where audio files have been grouped based on their original Emo Soundscapes DB emotion rating (i. e. High valence/arousal, and Low valence/arousal) – we see that although consistently different to the original source, High emotion does remain to higher than low emotional audio groups. Suggesting that trends in the audio files which are inherent to the emotion are left unchanged. However, this assumption requires further study.

Given this naive approach, further adaptation and audio choices based on emotional content may see further improvements in affective change. It is also worth noting that the audio applied here is extremely rudimentary, and further digital signal processing techniques, along with the use of more typically 'pleasing' audio may would be of value to explore.

4 A Deeper Approach for Soundscape Augmentation

Based on our initial findings, in this section we briefly outline a methodology for a soundscape augmentation, which is based on an individual's current state, and would be applied in further studies by the authors on this topic. An overview of this system is given in Fig. 3. Predominately an audio-based approach, we aim to utilise methodologies from the field of Speech Emotion Recognition (SER) [54], as well Generative Voice Conversion [28]. First in this section, we outline the feature extraction method for understanding and individuals state. Following this an *offline system*, in which the user would define a duration of listening, in a quiet space is defined. We also propose an *online system*, which in real-time 'augments' the natural soundscape, through sonification of audio features, generated based on emotional understanding of the user.

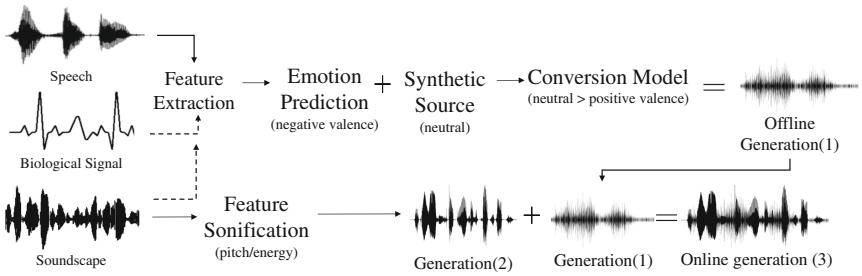

Fig. 3. Overview of the proposed affective soundscape generation system for wellbeing, via multimodal monitoring.

4.1 Feature Extraction and Emotional Prediction

From the user's input, a fusion of features known to the affective computing community (e. g., MFCCs, and spectral) [29], can be extracted from multiple modes (including the voice, as well as the current soundscape). Of note, in recent works we have found a correlation between biological signals, including hormone-based cortisol and speech features, during a stressful situation [3], suggesting that handcrafted features may be useful in this context to gain an understanding of a user state of lower wellbeing. As well this, if appropriate based on user-device, biological feature can also be utilised for understanding states of lower wellbeing [2].

Utilising deep, pre-trained neural networks, the tailored feature sets can then be classified for their emotionality (e. g., level of arousal and valence) [32]. The resulting, prediction are then used to define the current state of a user, as a condition for audio generation.

4.2 Offline Audio Generation

For the offline generation, perhaps in the scenario where a listener aims to reduce their affective state for short-term period, a pre-existing synthetic emotional source could be used. In this case, a dataset of synthetic audio could be applied, such as the richly annotated EmoSynth database [5]. From this, one-minute emotional samples can be created based on their emotional values; typically, this equates to aspects in audio such as, high arousal being equal to higher pitch and low arousal being equal lower pitch, with valence being a somewhat more complex aspect of emotion in terms of acoustic representation. As a means of obtaining varied (i. e. novel for each user interaction) audio outputs for each user, with more fine-grained differences, a generative adversarial strategy can be applied, such as 'StarGAN' [28]. In this scenario, a network can be trained on a selection of emotional classes. Following this depending on the given emotional prediction (or target) of the individuals state, a synthetic soundscape (the source) is then generated based on the target (user defined) emotion,e. g. if the user is in a state of high arousal, a low aroused soundscape is generated, for a (user-defined) given period of listening.

4.3 Online Generation Including Feature Sonification

For longer interaction periods, possibly even continual (i. e. for implementation in a chaotic working environment), we propose a method in which the offline audio generation is combined with a sonification of the features from the natural ongoing soundscape. To summarise this process, features such as, *chromatic, energy, and F0* can be extracted from the incoming soundscape signal, and reasoning be applied to sonify the Chroma and pitch-based features based on the energy of the signal at a given time-point. As well as this, the natural rhythm of the soundscape can be extracted and as an option, then applied to the resulting real-time generation. Rhythm is included, as a consistent rhythm has shown to have positive effects on wellbeing, producing a calming affect [36]. These two sonification approaches (feature-based and rhythm) are then applied to the offline generation process previously described, and the user is able to balance the level for each.

5 Conclusion and Outlook

In this contribution, we made preliminary user studies on the effect of augmenting natural soundscapes, as well as proposing a 'next-step' methodology for a personalised version of such a system. A series of *perception studies* [39] including those by the authors [5], support the initial assumption that specific combinations of audio can alter states of individual wellbeing - and initial results in this contribution also show similar trends. Thus, these findings support further development of the work described herein.

When *monitoring states* of poor wellbeing, there are many emotional states linked to this, prior work by the authors has focused on public-facing speech, as a marker of stress [3]. Findings have shown that through the use of a combination of conventional acoustic features, and machine learning algorithms, biological signals including skin conductance, blood volume pressure, and cortisol can be predicted during such states of lower wellbeing. Based on this, it would be of great interest to approach the development of a multimodal system, however with audio monitoring being non-intrusive and lower in resources, it may alone be the optimal modality.

In regards to audio generation, a deep auto regressive generative model such as WaveNet [41] has shown promise for generating affective data [4], and through the use of a generative adversarial network, the authors are currently experimenting with emotional data in a conversion paradigm, i.e., from one emotion to the other, e. g., happy to sad. Integrating such a generation method here, may allow for more variety in generation, however a naive training approach based on single emotions does also show promise for the desired outcome.

Acknowledgements. This work is funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

References

1. Amiriparian, S., Freitag, M., Cummins, N., Gerzcuk, M., Pugachevskiy, S., Schuller, B.W.: A fusion of deep convolutional generative adversarial networks and sequence to sequence autoencoders for acoustic scene classification. In: Proceedings of 26th European Signal Processing Conference (EUSIPCO), EURASIP, pp. 982–986. IEEE, Rome (2018)
2. Baird, A., Amiriparian, S., Berschneider, M., Schmitt, M., Schuller, B.: Predicting blood volume pulse and skin conductance from speech: introducing a novel database and results. In: Proceedings IEEE 21st International Workshop on Multimedia Signal Processing, MMSP 2019, 5 pages. IEEE, Kuala Lumpur, September 2019
3. Baird, A., et al.: Using speech to predict sequentially measured cortisol levels during a trier social stress test. In: Proceedings Interspeech 2019, pp. 534–538 (2019)
4. Baird, A., Amiriparian, S., Schuller, B.: Can deep generative audio be emotional? Towards an approach for personalised emotional audio generation. In: Proceedings IEEE 21st International Workshop on Multimedia Signal Processing, MMSP 2019, 5 pages. IEEE, Kuala Lumpur, September 2019
5. Baird, A., Parada-Cabaleiro, E., Fraser, C., Hantke, S., Schuller, B.: The perceived emotion of isolated synthetic audio: the emosynth dataset and results. In: Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion, p. 7. ACM (2018)
6. Baird, A., Schuller, B.: Acoustic sounds for wellbeing: a novel dataset and baseline results (2019)
7. Baumgartner, T., Esslen, M., Jäncke, L.: From emotion perception to emotion experience: emotions evoked by pictures and classical music. Int. J. Psychophysiol. **60**(1), 34–43 (2006)
8. Blood, A.J., Zatorre, R.J., Bermudez, P., Evans, A.C.: Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions. Nat. Neurosci. **2**(4), 382 (1999)
9. Brown, P.L.: In the classroom, a new focus on quieting the mind. https://www.nytimes.com/2007/06/16/us/16mindful.html. Accessed 2 Feb 2019
10. Burraston, D., Edmonds, E., Livingston, D., Miranda, E.R.: Cellular automata in midi based computer music. In: Proceedings of the 2004 International Computer Music Conference, p. no pagination. International Computer Music Association (2004)
11. Calvo, R.A., D'Mello, S., Gratch, J.M., Kappas, A.: The Oxford Handbook of Affective Computing. Oxford University Press, Oxford (2015)
12. Chu, S., Narayanan, S., Kuo, C.C.J.: Environmental sound recognition with time-frequency audio features. IEEE Trans. Audio, Speech Lang. Process. **17**(6), 1142–1158 (2009)
13. Dijk, E.O., Nijholt, A., Van Erp, J.B., Van Wolferen, G., Kuyper, E.: Audio-tactile stimulation: a tool to improve health and well-being? Int. J. Auton. Adap. Commun. Syst. **6**(4), 305–323 (2013)
14. Donahue, C., McAuley, J., Puckette, M.: Synthesizing audio with generative adversarial networks. CoRR abs/1802.04208 (2018)
15. Drossos, K., Floros, A., Kanellopoulos, N.G.: Affective acoustic ecology: towards emotionally enhanced sound events. In: Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound, pp. 109–116. ACM (2012)

16. Engel, J., et al.: Neural audio synthesis of musical notes with wavenet autoencoders. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1068–1077. JMLR. org (2017)
17. Fan, J., Thorogood, M., Pasquier, P.: Emo-soundscapes: A dataset for soundscape emotion recognition. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 196–201. IEEE (2017)
18. Fried, R.: Integrating music in breathing training and relaxation: II. applications. Biofeedback Self-regul. **15**(2), 936–943 (1990). (171–177)
19. Frigeni, B., et al.: Chemotherapy-induced peripheral neurotoxicity can be misdiagnosed by the national cancer institute common toxicity scale. J. Peripheral Nerv. Syst. **16**(3), 228–236 (2011)
20. Goldsby, T.L., Goldsby, M.E., McWalters, M., Mills, P.J.: Effects of singing bowl sound meditation on mood, tension, and well-being: an observational study. J. Evid.-Based Complement. Altern. Med. **22**(3), 401–406 (2017)
21. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
22. Handel, M.J.: mhealth (mobile health)–using apps for health and wellness. Explore **7**(4), 256–261 (2011)
23. Hartmann, K., Siegert, I., Philippou-Hübner, D., Wendemuth, A.: Emotion detection in HCI: from speech features to emotion space. IFAC Proc. Vol. **46**(15), 288–295 (2013)
24. Humphries, K.: Healing Sound: Contemporary Methods for Tibetan Singing Bowls. Ph.D. thesis, Loyola Marymount University, CA, US (2010)
25. Iyendo, T.O.: Exploring the effect of sound and music on health in hospital settings: a narrative review. Int. J. Nurs. Stud. **63**, 82–100 (2016)
26. Juslin, P.N., Västfjäll, D.: Emotional responses to music: the need to consider underlying mechanisms. Behav. Brain Sci. **31**(5), 559–575 (2008)
27. Kabat-Zinn, J., et al.: Effectiveness of a meditation-based stress reduction program. J. Psychiatry **149**(7), 936–943 (1992)
28. Kameoka, H., Kaneko, T., Tanaka, K., Hojo, N.: StarGAN-VC: non-parallel many-to-many voice conversion using star generative adversarial networks. In: 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 266–273. IEEE (2018)
29. Kishore, K.K., Satish, P.K.: Emotion recognition in speech using MFCC and wavelet features. In: 2013 3rd IEEE International Advance Computing Conference (IACC), pp. 842–847. IEEE (2013)
30. Kortchmar, L., Vorländer, M., Slama, J.: Sound quality evaluation for the workplace: research on the influence of spatial sound distributions. Acta Acust. United Acust. **87**(4), 495–499 (2001)
31. Krichagin, V.: Health effects of noise exposure. J. Sound Vibr. **59**(1), 65–71 (1978)
32. Lalitha, S., Geyasruti, D., Narayanan, R., Shravani, M.: Emotion detection using MFCC and cepstrum features. Proc. Comput. Sci. **70**, 29–35 (2015)
33. Lee, C.Y., Toffy, A., Jung, G.J., Han, W.: Conditional wavegan. CoRR abs/1809.10636 (2018)
34. Lundqvist, L.O., Carlsson, F., Hilmersson, P., Juslin, P.N.: Emotional responses to music: experience, expression, and physiology. Psychol. Music **37**(1), 61–90 (2009)
35. Manzelli, R., Thakkar, V., Siahkamari, A., Kulis, B.: Conditioning deep generative raw audio models for structured automatic music. arXiv preprint arXiv:1806.09905 (2018)
36. Maurer, R.L., Kumar, V., Woodside, L., Pekala, R.J.: Phenomenological experience in response to monotonous drumming and hypnotizability. Am. J. Clin. Hypn. **40**(2), 130–145 (1997)

37. Mehri, S., et al.: Samplernn: An unconditional end-to-end neural audio generation model. arXiv preprint arXiv:1612.07837 (2016)
38. Mirelman, A., et al.: Audio-biofeedback training for posture and balance in patients with parkinson's disease. J. Neuroeng. Rehabil. **8**(1), 35 (2011)
39. Moscoso, P., Peck, M., Eldridge, A.: Systematic literature review on the association between soundscape and ecological/human wellbeing (2018)
40. Okcu, S., Ryherd, E.E., Zimring, C., Samuels, O.: Soundscape evaluations in two critical healthcare settings with different designs. J. Acoust. Soc. Am. **130**(3), 387–392 (2011)
41. van den Oord, A., et al.: Wavenet: A generative model for raw audio. CoRR abs/1609.03499, 4 (2016)
42. van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., Kavukcuoglu, K.: Conditional image generation with pixelcnn decoders. CoRR abs/1606.05328 (2016)
43. Panda, S.P.: Automated speech recognition system in advancement of human-computer interaction. In: 2017 International Conference on Computing Methodologies and Communication (ICCMC), pp. 302–306. IEEE (2017)
44. Parada-Cabaleiro, E., Baird, A.E., Cummins, N., Schuller, B.: Stimulation of psychological listener experiences by semi-automatically composed electroacoustic environments. In: Proceedings ICME 2017, pp. 1051–1056. IEEE, Hong Kong, July 2017
45. Paterson, N., Naliuka, K., Jensen, S.K., Carrigy, T., Haahr, M., Conway, F.: Design, implementation and evaluation of audio for a location aware augmented reality game. In: Proceedings of the 3rd International Conference on Fun and Games, pp. 149–156. ACM (2010)
46. Porcheron, M., Fischer, J.E., Reeves, S., Sharples, S.: Voice interfaces in everyday life. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2018)
47. Roden, T., Parberry, I.: Designing a narrative-based audio only 3D game engine. In: Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, pp. 274–277. ACM (2005)
48. Rogers, K., Nacke, L.E.: Exploring the potential of game audio for wellbeing. In: PGW@ CHI PLAY (2017)
49. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**(6), 1161 (1980)
50. Salamon, E., Kim, M., Beaulieu, J., Stefano, G.B.: Sound therapy induced relaxation: down regulating stress processes and pathologies. Med. Sci. Monitor **9**(5), 96–100 (2003)
51. Sanderson, P.: Cognitive work analysis and the analysis, design, and evaluation of human-computer interactive systems. In: Proceedings 1998 Australasian Computer Human Interaction Conference. OzCHI 1998 (Cat. No. 98EX234), pp. 220–227. IEEE (1998)
52. Schafer, R.M.: The Soundscape: Our Sonic Environment and the Tuning of the World. Inner Traditions Bear & Co., Vermont (1993)
53. Schirosa, M., Janer, J., Kersten, S., Roma, G.: A system for soundscape generation, composition and streaming. In: XVII CIM-Colloquium of Musical Informatics, p. no pagination (2010)
54. Schuller, B., Rigoll, G., Lang, M.: Hidden markov model-based speech emotion recognition. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP 2003), vol. 2, pp. II-1. IEEE (2003)

55. Sebe, N., Cohen, I., Gevers, T., Huang, T.S.: Emotion recognition based on joint visual and audio cues. In: 18th International Conference on Pattern Recognition (ICPR 2006), vol. 1, pp. 1136–1139. IEEE (2006)
56. Sliwinski, J., Katsikitis, M., Jones, C.M.: Mindful gaming: how digital games can improve mindfulness. In: Abascal, J., Barbosa, S., Fetter, M., Gross, T., Palanque, P., Winckler, M. (eds.) INTERACT 2015. LNCS, vol. 9298, pp. 167–184. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22698-9_12
57. Van Den Oord, A., et al.: Wavenet: A generative model for raw audio. CoRR abs/1609.03499 (2016)
58. Williamson, I., Leeming, D., Lyttle, S., Johnson, S.: 'It should be the most natural thing in the world': exploring first-time mothers' breastfeeding difficulties in the UK using audio-diaries and interviews. Matern. Child Nutr. **8**(4), 434–447 (2012)
59. Wood, L., Ivery, P., Donovan, R., Lambin, E.: "To the beat of a different drum": improving the social and mental wellbeing of at-risk young people through drumming. J. Publ. Mental Health **12**(2), 70–79 (2013)
60. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
61. Zheng, A., et al.: Effects of a low-frequency sound wave therapy programme on functional capacity, blood circulation and bone metabolism in frail old men and women. Clin. Rehabil. **23**(10), 897–908 (2009)