ON THE SUITABILITY OF EVOLUTIONARY COMPUTING TO DEVELOPING TOOLS FOR INTELLIGENT MUSIC PRODUCTION

Alex Wilson
Acoustics Research Centre
University of Salford
Salford, UK
a.wilson1@edu.salford.ac.uk

Róisín Loughran
Natural Computing Research and
Applications Group,
University College Dublin
Dublin, Ireland
roisin.loughran@ucd.ie

Bruno M. Fazenda
Acoustics Research Centre
University of Salford
Salford, UK
b.m.fazenda@salford.ac.uk

ABSTRACT

Intelligent music production tools aim to assist the user by automating music production tasks. Many previous systems sought to create the best possible mix based on technical parameters but rarely has subjectivity been directly incorporated. This paper proposes that a new generation of tools can be designed based on evolutionary computation methods, which are particularly suited to dealing with the non-linearities and complex solution spaces introduced by perceptual evaluation. These techniques are well-suited to studio applications, in contrast to many previous systems which prioritized the live environment. Furthermore, there is potential to address accessibility issues in existing systems which rely greatly on visual feedback. A survey of previous literature is provided before the current state-of-the-art is described and a number of suggestions for future directions in the field are made.

1. INTRODUCTION

The art of mixing multitrack audio can be considered a complex act of multi-objective optimisation. The aim is to produce a mix that satisfies a number of criteria, many of which are highly subjective in nature. For example, in a macro sense, one may wish to maximise *"quality"*, which can be achieved by optimising other characteristics, such as *"clarity"*, *"warmth"*, *"punchiness"*, as well as conveying the desired emotional characteristics and artistic intent. Solving such a problem presents a number of possible issues, such as identifying a solution space, adapting to the geometry of this space and the possibly large number of parameters, and the subjective nature of audio quality. It is proposed in this paper that Evolutionary Computation (EC) can be used to address some of these challenges.

EC methods mimic the process of evolution by considering a population of individual solutions over a series of successive generations, rather than trying to deterministically improve one single solution. At the beginning of an EC run, a population of random solutions to the given problem is created. Each individual solution is assigned a *fitness* according to how well it solves that problem. The solutions are then selected for survival and reproduction into the next generation based on this fitness; solutions that are adept at solving the problem (have good fitness) are more likely to be selected than those that are not. These selected solutions are then modified using operator functions such as mutation or crossover in creating the subsequent population of solutions. As this process is repeated, the performance of the overall population of solutions is improved and the best-performing candidate in the final population can be chosen as the solution to the given problem.

EC methods have been developed from using a string-based genomic representation as used in Genetic Algorithms [1], to using tree based representations in Genetic Programming [2] or developing a genome-phenome grammar to map from a linear genome into a more useful representational domain in Grammatical Evolution [3]. These methods were developed using traditional problems that had a specific optimal solution such as symbolic regression and the artificial ant trail. The solutions to such problems have simple representations and are easily measured with a numerical fitness. In more recent years such methods have been applied to a variety of music-related applications such as composition [4], generating jazz solos [5] and musical instrument recognition [6].

2. SUITABILITY OF EVOLUTIONARY COMPUTING TO INTELLIGENT MUSIC PRODUCTION

This paper proposes that there exists a good argument as to why EC is well-suited to IMP problems. This argument is based on the following.

Non-linearities — Due to the perceptual nature of audio evaluation, the solution space may not be smooth and differentiable, making optimisation methods such as gradient descent difficult or impossible to apply. Additionally, as each user may have a different goal in mind, there may not exist a single global optimum. Each user may perceive a "*personal* global optimum" rather than every user agreeing on a "*universal* global optimum".

Large number of parameters — Often there are a large number of parameters where the relationships between them are not well-understood. Furthering the understanding of these relationships helps construct more efficient search spaces. It is also important to establish the mapping between system parameters and perceptual factors.

Fitness functions — The definition of a "good" mix, or at least a desired mix, can be complex but is ultimately subjective. What is required is a numerical value for fitness. Quantities to be minimised include the distance to a desired target which is known in advance, or quantities thought to degrade audio quality, such as inter-channel masking [7]. However, if perceptual targets are being sought, such as "warmth" or "clarity", explicit subjective ratings can be used as a fitness function in place of a numerical approximation.

A synthesis of these three observations leads to the use of Interactive Evolutionary Computing/Computation (IEC). IEC is a form of EC in which the fitness evaluation is not based on a clearly defined formula but on the subjective response of a user. IEC has been utilised in the solution of various problems which are subjective, such as fashion design [8], logo design [9] and sound synthesis [10] (see [11] for a detailed review of applications). In IEC, the system generates solutions in the problem's parameter space while the user evaluates the fitness of the solution in some psychological space, which may be unique to each user.

Notably, the above examples all incorporate design problems in which aesthetics are important. In such applications, there may not be a clearly defined optimal solution, that is considered suitable for a range of users. Neither is the fitness landscape clearly defined. The fitness function depends greatly on what is asked of the user conducting the evaluation and their understanding of the question posed and the domain of the problem. For example, in the case of fashion design, users may be asked to rate the fitness of presented candidate solutions (outfits) where the target is a series of descriptions such as *"warm", "smart", "casual", autumnal"* etc. and each user may have a different understanding of these concepts and how they are realised in an outfit.

Considering that a user must evaluate the fitness of each solution, this can become a time-consuming activity, with potential for high levels of cognitive demand and eventual fatigue. This is especially problematic in audio, where each individual solution may take tens of seconds to evaluate, compared to visual stimuli, where a number of solutions can be compared side-by-side. Of course, since the solutions are evaluated by audition only, there is no obvious need for visual displays to be used and this has advantages in terms of accessibility. To avoid user-fatigue, in parallel to the emergence of IEC has been the development of hybrid methods in which a relatively small number of solutions are evaluated by the user and the fitness of remaining solutions is merely inferred. This reduces the burden on the user for problem types where large populations are required. One such approach is to use clustering of solutions [12].

If "quality" is the variable to be optimised one must appreciate that quality can be considered as specific to a single product, good or service [13]. A framework for quality assessment has been provided suggesting that the concept of quality can be considered from four points of view [14]: quality as excellence or superiority, quality as value, quality as conforming to specifications, quality as meeting or exceeding customer expectations. While the third could possibly lead to an objective fitness function, the other perspectives suggest subjective evaluation. For example, listener expectations may differ by genre — an electric guitar may be processed differently in hip-hop compared to heavy metal. This subjectivity furthers the case for using IEC.

3. PREVIOUS WORK

Many prior works are based on matching a sound or mix to a target, using the distance from the target as a fitness function to be minimised. Of course, this target must be known in advance. Heise et al. [15] compared four techniques (including genetic algorithm and particle swarm optimisation) in the task of adjusting the parameters of a reverberation plug-in to best match a given room impulse response. Kolasinski [16] was concerned with matching a mix to a target, by adjusting track gains and using the Euclidean distance between spectral histograms as a similarity measure that was to be minimised using GA. Barchiesi & Reiss [17] also attempted matching to a given target mix, by optimising track gains and track EQ filters, using least-squares. This paper was critical of GA in general for this application, stating that *"... for the purpose of this application, the results are quite poor as the number of tracks increases and the algorithm is computationally expensive."* These performance issues may not have been due to high-dimensionality per se, but rather the choice of an inefficient solution space. It was later shown that defining the mix as the sum of input tracks produces a sub-optimal solution space for automated mixing and that optimisation of track gains and EQ filters benefits from carefully designed solution spaces, in which each possible configuration exists only once [18].

There are many more papers on various "matching to a target" applications [19, 20]. What about when there is no target audio available? In place of an explicit target audio there may still exist a target in some other domain, such as a perceptual target ("Make the mix sound bright/warm...etc").

4. FUTURE DIRECTIONS

For the problem of automated gain adjustment, a solution space has been proposed in the form of a series of inter-channel level balances between audio tracks [18]. This allows any mix of those tracks to be retrieved. The exploration of this space by means of an Interactive Genetic Algorithm (IGA) has been proposed where users consider a mix based solely on auditory perception and offer a rating between 1 and 10 using a numerical keypad [21]. Quantitative and qualitative evaluation indicates that the system is capable of producing desired mixes comparable to those produced by a traditional fader-based mixing interface presented on-screen, yet without the associated visual or physical demands [22].

It is possible for this space to be extended to include equalisation parameters and pan positions [22]. However, the number of variables to optimise can then become very large. This is where hierarchical structure in music mixing can be exploited, such as sub-grouping [23], by breaking down the problem into a series of smaller problems. Adding EQ to a kick drum, mixing the drum tracks, repeating for other instrument groups, mixing the subgroups — these are examples of relatively small tasks which, when combined, form a complete mixing session.

Memetic algorithms (MA) [24] could potentially improve performance. MA involves not only the evolution of a population by means of their genes but the more general evolution of a culture/society, which consists of both genetic and non-genetic evolution which can evolve in parallel. Non-genetic evolution involves the production and dissemination of "memes", which spread throughout the population. Those memes which prove useful to the population are adapted and passed on to subsequent generations. While the "genes" in [21] are the inter-channel balances between instruments, the "memes" in the population could be any of the following example strategies:

bright: mixes should sound "brighter", which can be achieved by higher spectral centroid

warm: mixes should sound "warmer", which can be achieved by lower spectral centroid

wide: mixes are considered better if they exhibit wide stereo impressions, achieved by panning and equalisation, and measured using audio signal features such as the stereo panning spectrogram [25].

punchy: preference for mixes that are punchier (having short periods of significant change in power), as determined by audio signal features [26].

This use of memes within the population allows certain assumptions to be placed into the system initially, such as "brighter mixes are better", only for the user to validate or reject these assumptions by their fitness ratings. Any specific quality can be introduced as a meme provided that quality can be measured or approximated from the mix. Once the genetic operations have taken place, memetic evolution can be implemented in a number of ways, such as performing heuristic-based local searches. Each individual solution has a probability of undergoing a local search, a meme (or set of memes) informing that search (i.e. search for mixes with higher/lower values of certain audio features).

From a previous study [27] we know, roughly, how these audio signal features are distributed in a population of mixes. The distribution of features such as spectral centroid, in mixes of a specific song, can be well-approximated by simple parametric models, such as a Gaussian distribution. A mix which is located far from the central tendency of such a distribution can be assumed to be an unnatural mix, one unlikely to be created by a real mix-engineer. The distance from this feature can therefore be used as a fitness-penalty.

5. CONCLUSIONS

From reviewing the literature we see that relatively early investigations into IMP did feature EC, yet did not manage to gain much traction during the recent revival in the topic. Perhaps this was due to the method's inherent "offline" nature, which may not be so well-suited to live applications as other deterministic, heuristic-based methods. EC methods can therefore be used in the development of novel studio-based technologies, particularly for automated mixing and various post-processing applications. A number of EC-based music production systems have been proposed and are currently under development. We propose a number of areas of study which have the potential to further our understanding of music mixing. More generally, it is hoped that the study of audio production will benefit from an increased emphasis on psychoacoustics and human subjectivity, as presented herein.

6. ACKNOWLEDGEMENTS

This work of the second author is part of the App'Ed (Applications of Evolutionary Design) project funded by Science Foundation Ireland under grant 13/IA/1850.

7. REFERENCES

[1] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," *Machine learning*, vol. 3, no. 2, pp. 95–99, 1988.

[2] J. R. Koza, *Genetic programming: on the programming of computers by means of natural selection*, vol. 1. MIT press, 1992.

[3] M. O'Neill and C. Ryan, "Grammatical evolution," *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 4, pp. 349–358, 2001.

[4] P. Dahlstedt, "Autonomous evolution of complete piano pieces and performances," in *Proceedings of the ECAL Workshop on Music and Artificial Life*, (Lisbon, Portugal), 2007.

[5] J. Biles, "Genjam: A genetic algorithm for generating jazz solos," in *Proceedings of the International Computer Music Conference*, pp. 131–131, International Computer Music Association, 1994.

[6] R. Loughran, J. Walker, M. O'Neill, and J. McDermott, "Genetic programming for musical sound analysis," in *International Conference on Evolutionary and Biologically Inspired Music and Art*, pp. 176–186, Springer, 2012.

[7] P. Aichinger, A. Sontacchi, and B. Schneider-Stickler, "Describing the transparency of mixdowns: The Masked-to-Unmasked-Ratio," in *130th AES Convention*, (London, UK), 2011.

[8] H.-S. Kim and S.-B. Cho, "Application of interactive genetic algorithm to fashion design," *Engineering Applications of Artificial Intelligence*, vol. 13, no. 6, pp. 635–644, 2000.

[9] M. O'Neill and A. Brabazon, "Evolving a logo design using Lindenmayer systems, postscript & grammatical evolution," in *IEEE World Congress on Computational Intelligence*, pp. 3788–3794, IEEE, 2008.

[10] J. McDermott, N. J. Griffith, and M. O'Neill, "Evolutionary GUIs for sound synthesis," in *Workshops on Applications of Evolutionary Computation*, pp. 547–556, Springer, 2007.

[11] H. Takagi, "Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation," *Proceedings of the IEEE*, vol. 89, no. 9, pp. 1275–1296, 2001.

[12] J.-Y. Lee and S.-B. Cho, "Sparse fitness evaluation for reducing user burden in interactive genetic algorithm," in *Fuzzy Systems Conference Proceedings*, vol. 2, pp. 998–1003, IEEE, 1999.

[13] E. Babakus and G. W. Boller, "An empirical assessment of the SERVQUAL scale," *Journal of Business Research*, vol. 24, no. 3, pp. 253–268, 1992.

[14] C. A. Reeves and D. A. Bednar, "Defining quality: Alternatives and implications.," *Academy of Management Review*, vol. 19, no. 3, pp. 419–445, 1994.

[15] S. Heise, M. Hlatky, and J. Loviscach, "Automatic adjustment of off-the-shelf reverberation effects," in *126th AES Convention*, (Munich, Germany), 2009.

[16] B. A. Kolasinski, "A framework for automatic mixing using timbral similarity measures and genetic optimization," in *124th AES Convention*, (Amsterdam, The Netherlands), pp. 1–8, 2008.

[17] D. Barchiesi and J. Reiss, "Automatic target mixing using least-squares optimization of gains and equalization settings," in *Proceedings of the 12th Conference on Digital Audio Effects (DAFx-09)*.

[18] A. Wilson and B. M. Fazenda, "Navigating the mix-space: theoretical and practical level-balancing technique in multitrack music mixtures," in *Proceedings of the Sound and Music Computing Conference*, (Maynooth, Ireland), July 2015.

[19] S. Ghosh, D. Kundu, K. Suresh, S. Das, and A. Abraham, "Design of optimal digital IIR filters by using a bandwidth adaptive harmony search algorithm," in *World Congress on Nature & Biologically Inspired Computing*, pp. 481–486, IEEE, 2009.

[20] M. F. Caetano and X. Rodet, "Independent manipulation of high-level spectral envelope shape features for sound morphing by means of evolutionary computation," in *International Conference on Digital Audio Effects (DAFx-10)*, (Graz, Austria), 2010.

[21] A. Wilson and B. Fazenda, "An evolutionary computation approach to intelligent music production, informed by experimentally gathered domain knowledge," in *2nd AES Workshop on Intelligent Music Production*, September 2016.

[22] A. Wilson, *Evaluation and modelling of perceived audio quality in popular music, towards intelligent music production*. PhD thesis, University of Salford, 2017.

[23] D. Ronan, H. Gunes, D. Moffat, and J. D. Reiss, "Automatic subgrouping of multitrack audio," in *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, (Trondheim, Norway), 2015.

[24] P. Moscato *et al.*, "On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms," *Caltech concurrent computation program, C3P Report*, 1989.

[25] G. Tzanetakis, R. Jones, and K. McNally, "Stereo panning features for classifying recording production style.," in *8th International Society for Music Information Retrieval Conference*, (Vienna, Austria), 2007.

[26] S. Fenton and H. Lee, "Towards a perceptual model of 'punch' in musical signals," in *139th AES Convention*, (New York, USA), 2015.

[27] A. Wilson and B. Fazenda, "Variation in multitrack mixes: Analysis of low-level audio signal features," *J. Audio Eng. Soc*, vol. 64, no. 7/8, pp. 466–473, 2016.